

# A proposed semantic keywords search engine for Indonesian Qur'an translation based on word embedding

Liza Trisnawati<sup>1,2</sup>, Noor Azah Binti Samsudin<sup>2</sup>, Shamsul Kamal Bin Ahmad Khalid<sup>2</sup>,  
Ezak Fadzrin Bin Ahmad Shaubari<sup>2</sup>, Sukri<sup>1,2</sup>, Zul Indra<sup>3</sup>

<sup>1</sup>Department of Informatics Engineering, Universitas Abdurrah, Pekanbaru, Indonesia

<sup>2</sup>Department of Software Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Malaysia

<sup>3</sup>Department of Computer Science, Universitas Riau, Pekanbaru, Indonesia

## Article Info

### Article history:

Received Feb 17, 2024

Revised Mar 11, 2024

Accepted Mar 28, 2024

### Keywords:

Al-Qur'an

Search engine

Semantic keywords

Word embedding

Word2vec

## ABSTRACT

Obtaining relevant information from the Holy Qur'an can be really challenging for people who cannot speak Arabic, such as the Indonesian people. One technology implementation which is commonly used to tackle this problem is to develop a search engine application for Al-Qur'an verses. This paper proposes a search engine based on semantic representation keywords for the Indonesian translation of the Al-Qur'an which consists of 3 phases i.e., data preparation, document representation, and search engine development. In the first stage, the Al-Qur'an dataset was built using the official translation of the Al-Qur'an from the Ministry of Religion and then enriched with the Wikipedia corpus. The second phase is document representation which produces feature vectors by utilizing the Word2Vec algorithm. Finally, the development of a search engine that can find the most relevant verses by calculating the cosine similarity between the document and the keywords. It was found that the proposed search engine succeeded in exceeding the performance of ordinary search engines by finding wider information due to the use of semantic keywords. Apart from that, the proposed search engine succeeded in maintaining the relevance of search results by achieving precision and recall levels of 98.7% and 97.3% respectively.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Noor Azah Samsudin

Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM)

Parit Raja, Batu Pahat, Malaysia

Email: azah@uthm.edu.my

## 1. INTRODUCTION

The Holy Qur'an is one of the most important things for Muslims in carrying out their lives as good Muslims. In the view of Islam, if someone is willing to be a good Muslim, they must be able to understand the contents of the Qur'an and underlie all of their activity to the guidance that is in the Qur'an. This is because the Qur'an is one of the two main sources of law which is a guide for Muslims in living their lives. Qur'an contains things that are related to faith, science, law, muamalah (rules governing behavior and procedures for human life), stories of people before, worship and, tajwid. As a document, the Holy Qur'an is a scripture that is very long and rich in content where it consists of 30 juz (part), 114 surahs, and 6236 verses and is written in Arabic. This can be really challenging for most Muslims who are not proficient in Arabic, for example, Indonesian Muslims who speak in Indonesian language, to understand it. It will require a lot of time for them to find the appropriate information in the Qur'an if they have to refer to the Qur'an for certain problems. Hence, developing technology to make it easier for Muslims to discover knowledge contained in

the Qur'an is one of the efforts that is worth to be done [1]. One of the technology implementations that is commonly used to assist Qur'an knowledge discovery is developing a search engine application for Al-Qur'an verses [2].

Thus far, a lot of research has been carried out to help search for information in the Al-Qur'an by developing search engine applications. In the last 5 years, it was found that several related studies had worked on this topic. Even though a lot of research has been carried out, there is a big gap in existing research regarding the inability of the developed search application to understand and search for implicit (contextual) information in the Al-Qur'an. The majority of existing search engines only apply the concept of labeling [3], indexing and ranking as well as question answering textually without trying to understand verses contextually [2], [4]-[6]. This contextual understanding is very important in understanding the information contained in it since the Holy Al-Qur'an contains many implied meanings which is the biggest challenge in finding relevant information [7], [8].

One solution that can be applied to tackle this issue is to represent the Qur'anic text by paying attention to the semantic relationships that exist in each verse. Based on an in-depth study of the literature, it was found that only a few studies had tried this approach [9]-[11]. Moreover, the application of text representation based on semantic relationships in case studies of translations of the Indonesian Al-Qur'an can be said to be very rarely discussed. Existing research only applies a regular keyword-based approach [12] or keywords enriched with a Glossary [13]. Only a few studies have applied semantic-based document representation as carried out by Purnama *et al.* [14]. However, this research has not implemented dataset enrichment during document representation phase in order to obtain semantic keywords. To be concluded, the summary of the research gap and contribution offered by this research is illustrated in Figure 1.

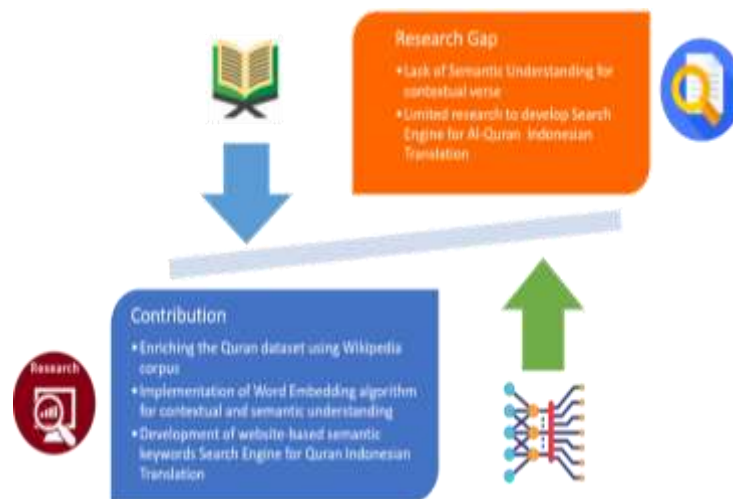


Figure 1. Summary of research gap and research contribution

Based on Figure 1, it is hoped that this research can bridge those research gaps by developing a semantic search engine application. This proposed application is expected to be able to understand semantic relationships to understand Qur'anic verses textually and contextually by generating semantic keywords. In particular, this search engine application is intended for Indonesian translations of Qur'anic verses because Indonesians have limitations in understanding Arabic documents which is an absolute requirement to obtain information from the Al-Qur'an.

## 2. METHOD

As stated in the introduction section, this study aims to facilitate the search for knowledge in the Al-Qur'an which contains a lot of implied information. However, this study discovered that there are three main challenges to achieve this goal (1) how to prepare the data, (2) how to represent data in semantic relationship, and (3) how to develop proposed semantic search engine for Al-Qur'an. Therefore, this study is divided into three stages to overcome these challenges i.e., data preparation, creating semantic relationships, and developing semantic search engine applications. The overall architecture of the research method to conduct this study is illustrated in Figure 2.

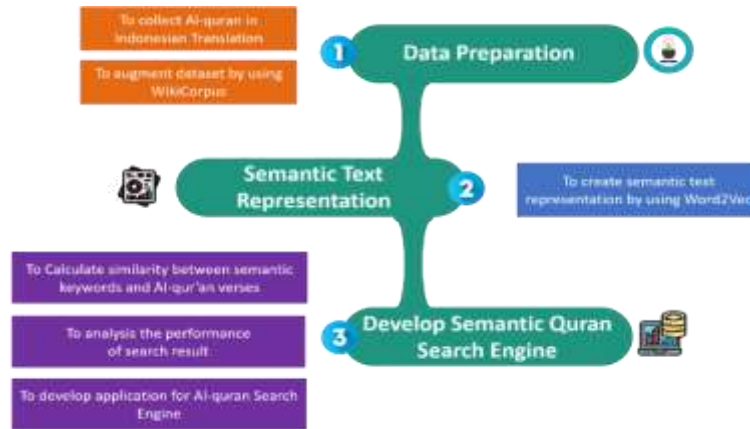


Figure 2. Research framework

### 2.1. Data preparation stage

As in general research related to data processing and information retrieval, the first stage in this research is data collection and preparation. As mentioned previously, the first problem in this research is how to prepare quality data for the Indonesian translation of the Al-Qur'an. To answer this problem, this research uses a translation of the Al-Qur'an sourced from the Indonesian Ministry of Religion. This translation can be considered the official and most valid translation of the Indonesian Al-Qur'an. Once the process of collecting the Indonesian Al-Qur'an dataset is conducted completely, the next task in the data preparation stage is augmenting the dataset. Enriching this dataset is very crucial for the next stage due to creating semantic relationships between texts requires a very large number of datasets.

### 2.2. Semantic text representation stage

The next stage of this research is the representation of Al-Qur'an verses based on their semantic relationship for the Al-Qur'an dataset that has been enriched in the first stage. Semantic relationships in a text refer to how words or phrases in a text interact meaningfully [15], [16]. Furthermore, these semantic relationships include the way words are related to each other and how the meaning of one part of the text can influence or be related to the meaning of other parts. In the field of natural language processing studies, especially in efforts to create a search engine for Al-Qur'an verses which contain many implied meanings, understanding semantic relationships is very important. Through this semantic relationship, a deeper understanding of the documents which are examined can be obtained [17]-[19], such as understanding implied meaning (contextual understanding), analogical reasoning, and increasing similarity calculations.

One method that is widely recommended for finding semantic relationships between words is to apply a word embedding algorithm [20], [21]. Semantic relationships in word embeddings refer to the way words are related or connected in meaning in a given vector space. This embedding captures semantic relationships by placing words with similar meanings close to each other in this space. Many algorithm choices can be used to represent documents based on their semantic relationships, such as Word2Vec, GloVe, or FastText [22]-[25]. In this research, the Word2Vec algorithm was chosen because of its ability to carry out analogous operations and understand contextual meaning better [26]-[28]. Apart from that, Word2Vec has advantages in its data processing capabilities where this algorithm is very appropriate for handling large data. The Word2Vec algorithm generally consists of 7 main stages i.e., choose Word2Vec model, prepare text corpus, tokenization and vocabulary formation, initialize word vectors, train the Word2Vec model, iterate through corpus and, update word vectors.

### 2.3. Search engine application development stage

The final stage of this research is the development of proposed search engine application for Indonesian translations of the Qur'an. Through this proposed application, it is hoped that users will be easier to find the information they need by simply entering the appropriate keywords. Furthermore, this proposed Qur'an search engine is a website-based application by employing the Django framework. Django is a Python-based web development framework designed to make it easy to create fast and efficient web applications [29]-[31]. Django is a popular choice for Python-based web development due to its ability to build small to large applications while maintaining good performance and high scalability [32]. The overall design for developing a search engine for the Al-Qur'an is shown in Figure 3.

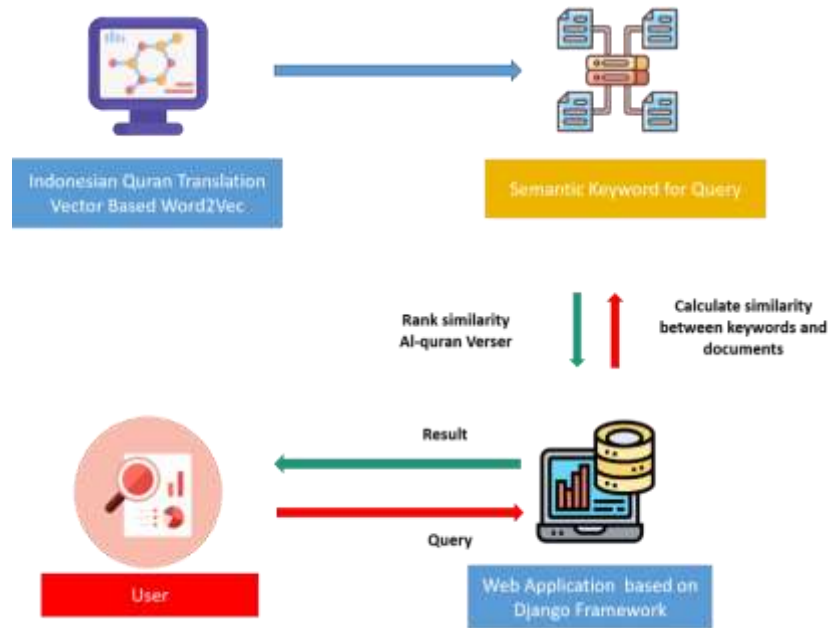


Figure 3. Flowchart for application development of AI-Qur'an search engine

Based on the design of Figure 3, it can be seen that this research uses the cosine similarity algorithm to measure the level of similarity between the semantic keyword and the AI-Qur'an word vector. Semantic keywords are obtained from expanding queries typed by users based on their semantic relationship with the enriched Qur'an dataset cosine similarity is a metric used to measure the extent to which two direction vectors are close or the extent to which two documents or data sets are similar in a high-dimensional vector space. Cosine similarity is measured as the cosine of the angle between two vectors [33]-[35]. For example, there are two vectors, namely vector  $X$  and vector  $Y$ , which are two vectors in  $n$ -dimensional vector space, hence the cosine similarity value ( $\cos \theta$ ) can be calculated using in (1):

$$\text{Cosine Similarity } (X, Y) = \frac{X \cdot Y}{|X| \cdot |Y|} \quad (1)$$

As for:

- $X \cdot Y$  is the result of the dot product between vectors  $X$  and  $Y$
- $|X|$  and  $|Y|$  is the Euclidean norm of each vector

Once the results of the cosine similarity calculation have been successfully obtained, the search engine will rank it based on the level of similarity and then display it to the user through the search engine application interface.

### 3. RESULTS AND DISCUSSION

The performance of the proposed semantic Qur'an search engine application is determined based on precision and recall measurements. By using several keywords, experiments were carried out on the semantic Qur'an search engine application which implemented semantic word embedding and the search engine application without word embedding (ordinary search engine). After that, a performance comparison will be carried out to find out whether the proposed semantic Qur'an search engine application can exceed the performance of ordinary search engine applications.

#### 3.1. Evaluation metrics

As previously mentioned, the performance of the AI-Qur'an verse search engine is calculated in terms of precision and recall. It is hoped that the proposed search engine can get high precision and recall values because this means that the search engine has good information search coverage as a sign that the search engine has succeeded in its expected performance. The measure of precision, also known as specificity, represents the correctness of information search results based on the ratio of relevant verses found (retrieved) to the total number of relevant and irrelevant verses found as shown in (2).

$$Precision = \frac{relevant\ verses \cap\ retrieved\ verses}{retrieved\ verses} \tag{2}$$

The recall measure, also known as sensitivity, shows performance based on the ratio of the number of relevant verses found to the number of all possible related verses as shown in (3).

$$Relevant = \frac{relevant\ verses \cap\ retrieved\ verses}{relevant\ verses} \tag{3}$$

**3.2. Discussion**

This research chose 7 topics i.e., prayer, zakat, fasting, pilgrimage, prophets, holy books, angels, apocalypse, trade, and clothing which are very important for Muslims as queries that will be tested. Augmented Al-Qur’an queries and datasets are converted into vectors based on semantic relationships using the word embedding Word2Vec algorithm. This semantic relationship-based vector conversion process will produce several keywords that will be used to search for information on Qur’anic verses according to relevant topics as described in Table 1 and Figure 4.

Table 1. Generated semantic keywords for topic

Query	Top 7 generated semantic keywords						
prayer	shalat/ 0.915	salat/ 0.885	berjamaah/ 0.799	solat/ 0.77	subuh/ 0.769	ashar/ 0.763	tarawih/ 0.755
zakat	infaq/ 0.768	infak/ 0.702	wakaf/ 0.692	shadaqah/ 0.669	amil/ 0.628	fitrah/ 0.603	pembiayaan/ 0.585
fasting	berpuasa/ 0.794	ramadan/ 0.722	berbuka/ 0.717	sholat/ 0.686	shalat/ 0.683	adha/ 0.668	maulud/ 0.646
pilgrimage	umrah/ 0.678	khatib/ 0.594	syaikhon/ 0.59	manasik/ 0.563	syech/ 0.552	arsyad/ 0.551	husin/ 0.549
prophets	nabi/ 0.657	injil/ 0.631	paulus/ 0.63	timotius/ 0.627	matus/ 0.616	yohanes/ 0.615	bapa/ 0.592
angels	jibril/0.667	israfil/ 0.617	iblis/ 0.598	tuhan/ 0.584	rasul/ 0.582	mikail/ 0.574	roh/ 0.572
apocalypse	berbangkit/ 0.555	penghakiman/ 0.543	qiyamat/ 0.52	petang/ 0.517	ajal/ 0.51	dajjal/ 0.508	harmagedon/ 0.505

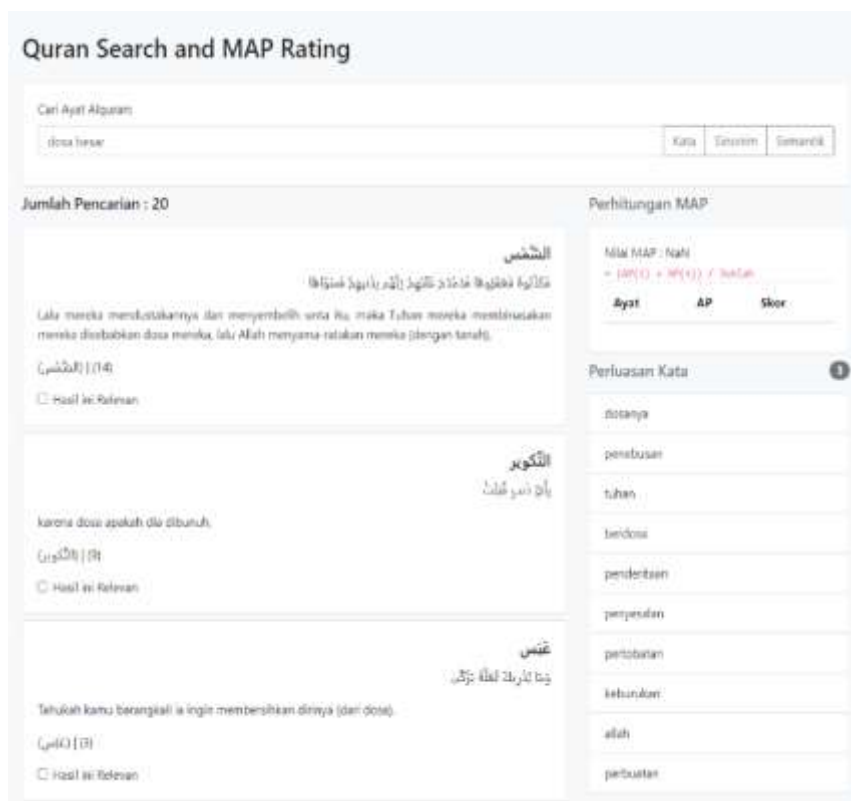


Figure 4. Search engine interface for Zakat topics

As seen in Table 1, each topic has several keywords based on their semantic relationships. It can be concluded that broader information can be obtained due to the large number of keywords that are formed instead of just relying on limited keywords as is done by ordinary search engines. Once the vector conversion process and keyword evocation are complete, the similarity calculation between the topic and the verses of the Al-Qur'an is carried out using the cosine similarity algorithm. Furthermore, sorting is carried out based on the results of this calculation to obtain the most relevant verses (keywords) for each topic as shown in Table 2.

Table 2. Comparison results for proposed semantic and ordinary search engine

Topic	Ordinary search engine		Proposed semantic search engine	
	Top 15 Qur'an verses	Relevant	Top 15 Qur'an verses	Relevant
Prayer	2/3; 2/43; 2/45; 2/83; 2/110; 2/125; 2/153; 2/177; 2/238; 2/239; 2/277; 3/39; 4/43; 4/102; 4/103;	56	108/2; 107/4; 98/5; 96/10; 75/31; 74/43; 70/22; 62/10; 62/9; 58/13; 42/38; 35/29; 33/33; 31/17; 31/4;	175
Zakat	2/43; 2/83; 2/110; 2/177; 2/277; 4/77; 4/162; 5/12; 5/55; 7/156; 9/5; 9/11; 9/18; 9/58; 9/60;	34	98/5; 73/20; 58/13; 41/7; 33/33; 31/4; 30/39; 27/3; 24/56; 24/37; 23/4; 22/78; 22/41; 21/73; 19/55;	94
Fasting	2/183; 2/184; 2/185; 2/187; 2/196; 4/92; 5/89; 5/95; 19/26; 33/35; 58/4; 66/5; 92/21	13	5/89; 2/187; 66/5; 58/4; 33/35; 19/26; 5/95; 4/92; 2/196; 2/185; 2/184; 2/183; 9/108; 9/11; 9/5;	56
Pilgrimage	2/128; 2/158; 2/189; 2/196; 2/197; 2/200; 3/97; 5/1; 5/2; 8/26; 8/72; 8/74; 9/3; 9/19; 9/100; 9/117;	21	22/27; 9/19; 9/3; 5/2; 5/1; 3/97; 2/197; 2/196; 2/189; 2/158; 2/128; 95/3; 90/2; 90/1; 73/15;	78
Prophets	2/87; 2/98; 2/101; 2/108; 2/129; 2/143; 2/151; 2/214; 2/253; 2/279; 2/285; 3/32; 3/49;	339	98/2; 91/13; 75/32; 75/31; 73/16; 73/15; 72/27; 69/40; 69/10; 65/11; 64/12; 63/1; 62/2; 61/6; 60/1; 59/7;	721
Angels	2/30; 2/31; 2/34; 2/98; 2/102; 2/161; 2/177; 2/210; 2/248; 2/285; 3/18; 3/39; 3/42; 3/45; 3/80;	142	97/4; 96/18; 89/22; 81/21; 80/15; 78/38; 74/31; 74/30; 72/27; 69/17; 54/6; 53/27; 53/26; 50/41; 50/21; 50/18;	437
Apocalypse	2/85; 2/113; 2/148; 2/165; 2/174; 2/210; 2/212; 3/55; 3/77; 3/161; 3/180; 3/185; 3/194; 4/87;	131	101/3; 101/2; 101/1; 79/34; 77/29; 75/6; 75/1; 71/18; 69/15; 69/4; 69/3; 69/2; 69/1; 68/39; 67/27;	395
Precision = 97.9%; Recall = 96.5%			Precision = 98.7%; Recall = 97.3%	

Based on the results displayed in Table 2, it can be concluded that the proposed semantic search engine has many advantages over ordinary search engines in terms of its ability to search for broader information and the relevance of search results. It can be seen that the proposed search engine can search for more information due to the expansion of queries based on semantic keywords obtained from the word embedding representation. By utilizing generated semantic keywords, this proposed search engine can obtain additional information such as adding synonyms of words and other implied meanings. Furthermore, even though the proposed search engine retrieves more information, it also manages to maintain the relevance of search results in terms of precision and recall levels. The proposed search engine succeeded in getting very good performance with precision and recall values of 98.7% and 97.3% respectively. Therefore, this proposed search engine is very reliable in helping search for information from Indonesian translations of the Al-Qur'an.

Even though it has been proven to be successful in exceeding the performance of ordinary search engines, the proposed search engine still has several weaknesses in terms of the keywords it produces. Sometimes the resulting semantic keywords are not related to the topic the user wants to search for, such as the topic of prophets. Several keywords from this topic are more suitable for Christian religious documents, such as the keywords Paul, and Timothy. This is due to the use of a wiki corpus which contains general words and is not specific to Islamic religious documents only. One solution that can be applied to overcome this problem is to add the concept of ontology to the semantic representation. Ontologies are expected to play a key role in searching for more specific information for certain domains. This is because ontology-based representation can enable users to search, organize, and present information in a more contextual and meaningful way.

#### 4. CONCLUSION

The research proposes a semantic representation-based search tool for Indonesian translations of the Al-Qur'an. This Al-Qur'an verse is then augmented with the Wiki corpus to get a richer dataset to produce better semantic relationships. Then by using the word embedding concept based on the Word2Vec algorithm,



this augmented dataset will be converted into a vector based on semantic relationships. Finally using the cosine similarity algorithm, the cosine similarity is calculated between the augmented dataset and the selected topic. The performance of the proposed search engine is measured by comparing its ability to a regular search engine (without word embedding and augmented dataset) in finding relevant verses based on precision and recall metrics. In addition, the relevance of the search results for Al-Qur'an verses was evaluated by Al-Qur'an experts as additional validation. The performance obtained was 98.7% and 97.3% for precision and recall. The next research that is worth doing is the application of ontology so that users can search, organize, and present information in a more contextual and meaningful way. Apart from that, implementing other words embedding algorithms such as GloVe and FastText is also interesting to work on.




## REFERENCES

- [1] D. I. A. Putra and M. Hidayatullah, "The roles of technology in al-Quran exegesis in Indonesia," *Technology in Society*, vol. 63, p. 101418, Nov. 2020, doi: 10.1016/j.techsoc.2020.101418.
- [2] Z. Indra, A. Adnan, and R. Salambue, "A hybrid information retrieval for Indonesian translation of Quran by using single pass clustering algorithm," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, IEEE, Oct. 2019, pp. 1–5. doi: 10.1109/ICIC47613.2019.8985737.
- [3] A. Adeleke, N. A. Samsudin, M. H. A. Rahim, S. K. A. Khalid, and R. Efendi, "Multi-label classification approach for quranic verses labeling," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 24, no. 1, p. 484, Oct. 2021, doi: 10.11591/ijeecs.v24.i1.pp484-490.
- [4] F. Beirade, "Search engine for Holy Quran," in *2014 4th International Symposium ISKO-Maghreb: Concepts and Tools for Knowledge Management (ISKO-Maghreb)*, IEEE, Nov. 2014, pp. 1–6. doi: 10.1109/ISKO-Maghreb.2014.7033477.
- [5] S. K. Hamed and M. J. A. Aziz, "A question answering system on Holy Quran translation based on question expansion technique and neural network classification," *Journal of Computer Science*, vol. 12, no. 3, pp. 169–177, Mar. 2016, doi: 10.3844/jcssp.2016.169.177.
- [6] R. H. Gusmita, Y. Durachman, S. Harun, A. F. Firmansyah, H. T. Sukmana, and A. Suhaimi, "A rule-based question answering system on relevant documents of Indonesian Quran translation," in *2014 International Conference on Cyber and IT Service Management (CITSM)*, IEEE, Nov. 2014, pp. 104–107. doi: 10.1109/CITSM.2014.7042185.
- [7] R. Z. Wan-chik, "Islamic and Quranic information on the web : information retrieval challenges and user' s preferences Islamic and Quranic information on the Web : Information," in *2nd International Conference on Islamic Applications in Computer Science And Technology*, Amman, Jordan, 2014, pp. 1–8.
- [8] Y. M. Alginahi, "Quran search engines: challenges and design requirements," *International Journal of Computer Applications in Technology*, vol. 57, no. 3, p. 237, 2018, doi: 10.1504/IJCAT.2018.092982.
- [9] M. I. E. K. Ghembaza, "Specialized Quranic semantic search engine," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 17, no. 2, 2019.
- [10] A. Hakkoum and S. Raghay, "Advanced search in the Qur'an using semantic modeling," in *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, IEEE, Nov. 2015, pp. 1–4. doi: 10.1109/AICCSA.2015.7507259.
- [11] E. H. Mohamed and E. M. Shokry, "QSST: a Quranic semantic search tool based on word embedding," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 3, pp. 934–945, 2022, doi: 10.1016/j.jksuci.2020.01.004.
- [12] E. M. A. Fenti, N. S. Rina, and A. Syukri, "Development of Qur'an Search engine for the Indonesian language query," in *Proceedings of the Proceedings of the 2nd International Conference on Quran and Hadith Studies Information Technology and Media in Conjunction with the 1st International Conference on Islam, Science and Technology, ICONQUHAS & ICONIST, Bandung, October 2-*, EAI, 2020. doi: 10.4108/eai.2-10-2018.2295579.
- [13] F. E. M. Agustin, M. H. R. Maulidi, R. H. Gusmita, R. C. N. Santi, M. Ulfa, and R. Sugara, "Applying of Quranic glossary approach to improve Indonesian Qur'an translation search engine performance," in *2020 8th International Conference on Cyber and IT Service Management (CITSM)*, IEEE, Oct. 2020, pp. 1–5. doi: 10.1109/CITSM50537.2020.9268820.
- [14] A. R. G.Purnama, I. N. Yulita, and A. Helen, "Search system for translation of Al-Qur'an verses in Indonesian using BM25 and semantic query expansion," in *2021 International Conference on Artificial Intelligence and Big Data Analytics*, IEEE, Oct. 2021, pp. 1–7. doi: 10.1109/ICAIBDA53487.2021.9689757.
- [15] B. Altnel and M. C. Ganiz, "Semantic text classification: a survey of past and recent advances," *Information Processing & Management*, vol. 54, no. 6, pp. 1129–1153, Nov. 2018, doi: 10.1016/j.ipm.2018.08.001.
- [16] R. A. Sinoara, J. Antunes, and S. O. Rezende, "Text mining and semantics: a systematic mapping study," *Journal of the Brazilian Computer Society*, vol. 23, no. 1, p. 9, Dec. 2017, doi: 10.1186/s13173-017-0058-7.
- [17] P. Yan and W. Jin, "Improving cross-document knowledge discovery using explicit semantic analysis," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, pp. 378–389. doi: 10.1007/978-3-642-32584-7\_31.
- [18] A. M. Rinaldi, C. Russo, and C. Tommasino, "A semantic approach for document classification using deep neural networks and multimedia knowledge graph," *Expert Systems with Applications*, vol. 169, p. 114320, May 2021, doi: 10.1016/j.eswa.2020.114320.
- [19] B. Aleman-Meza, "Searching and ranking documents based on semantic relationships," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, IEEE, 2006, pp. x136–x136. doi: 10.1109/ICDEW.2006.131.
- [20] Y. Li and T. Yang, "Word embedding for understanding natural language: a survey," in *Studies in Big Data*, 2018, pp. 83–104. doi: 10.1007/978-3-319-53817-4\_4.
- [21] Z. Yao, Y. Sun, W. Ding, N. Rao, and H. Xiong, "Dynamic word embeddings for evolving semantic discovery," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, New York, NY, USA: ACM, Feb. 2018, pp. 673–681. doi: 10.1145/3159652.3159703.
- [22] E. M. Dharna, F. L. Gaol, H. L. H. S. Warnars, and B. Soewito, "The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (CNN) text classification," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 2, pp. 349–359, 2022.
- [23] R. Egger, "Text representations and word embeddings," in *Tourism on the Verge*, 2022, pp. 335–361. doi: 10.1007/978-3-030-88389-8\_16.

- [24] Z. H. Kilimci and S. Akyokus, "The evaluation of word embedding models and deep learning algorithms for Turkish text classification," in *2019 4th International Conference on Computer Science and Engineering (UBMK)*, IEEE, Sep. 2019, pp. 548–553. doi: 10.1109/UBMK.2019.8907027.
- [25] R. A. Stein, P. A. Jaques, and J. F. Valiati, "An analysis of hierarchical text classification using word embeddings," *Information Sciences*, vol. 471, pp. 216–232, Jan. 2019, doi: 10.1016/j.ins.2018.09.001.
- [26] S. Yilmaz and S. Toklu, "A deep learning analysis on question classification task using Word2vec representations," *Neural Computing and Applications*, vol. 32, no. 7, pp. 2909–2928, Apr. 2020, doi: 10.1007/s00521-020-04725-w.
- [27] I. Skelac and A. Jandrić, "Meaning as use: from Wittgenstein to Google's Word2vec," in *Guide to Deep Learning Basics*, Cham: Springer International Publishing, 2020, pp. 41–53. doi: 10.1007/978-3-030-37591-1\_5.
- [28] S. Balli and O. Karasoy, "Development of content-based SMS classification application by using Word2Vec-based feature extraction," *IET Software*, vol. 13, no. 4, pp. 295–304, Aug. 2019, doi: 10.1049/iet-sen.2018.5046.
- [29] Gore *et al.*, "Django: web development simple & fast," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 6, pp. 4576–4585, 2021.
- [30] S. Chen, S. Ahmmed, K. Lal, and C. Deming, "Django web development framework: powering the modern web," *American Journal of Trade and Policy*, vol. 7, no. 3, pp. 99–106, Dec. 2020, doi: 10.18034/ajtp.v7i3.675.
- [31] P. Thakur and P. Jadon, "Django: developing web using Python," in *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, IEEE, May 2023, pp. 303–306. doi: 10.1109/ICACITE57410.2023.10183246.
- [32] K. Schutt and O. Balci, "Cloud software development platforms: a comparative overview," in *2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)*, IEEE, Jun. 2016, pp. 3–13. doi: 10.1109/SERA.2016.7516122.
- [33] J. Jaafar, Z. Indra, and N. Zamin, "A category classification algorithm for Indonesian and Malay news documents," *Jurnal Teknologi*, vol. 78, no. 8–2, Aug. 2016, doi: 10.11113/jt.v78.9549.
- [34] A. L. Jousselme and P. Maupin, "Distances in evidence theory: comprehensive survey and generalizations," *International Journal of Approximate Reasoning*, vol. 53, no. 2, pp. 118–145, Feb. 2012, doi: 10.1016/j.ijar.2011.07.006.
- [35] J. Wang and Y. Dong, "Measurement of text similarity: a survey," *Information*, vol. 11, no. 9, p. 421, Aug. 2020, doi: 10.3390/info11090421.

## BIOGRAPHIES OF AUTHORS






**Liza Trisnawati, S.T., M.Kom.**    is a lecturer at Abdurrah University, Indonesia. She obtained her Master in Information Technology from Putera Indonesia University (YPTK, Indonesia). He is currently pursuing his Ph.D. in Doctor of Philosophy of Information Technology at Tun Hussein Universiti Onn Malaysia (UTHM). Her research interests include data mining and machine learning. She can be contacted at email: liza.trisnawati@univrab.ac.id.







**Assoc. Prof. Dr. Noor Azah Binti Samsudin**    is Associate Professor at the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM). She holds a Ph.D. degree from the University of Queensland. Bachelor Of Computer Science, Universiti Kebangsaan Malaysia, Bachelor Of Computer Science, University Of Missouri, Columbia, US. She is Deputy Director of the Center for Academic Excellence and Development and Principal Investigator of the Faculty of Computer Science and Information Technology, Soft Computing, and Data Mining Center (SMC). Her areas of expertise are information, computer and communications technology (ICT), artificial intelligence, and machine learning. Her research area interests were then classification and feature selection. She can be contacted at email: azah@uthm.edu.my.







**Dr. Shamsul Kamal Bin Ahmad Khalid**    is lecturer at the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM). The B.Sc. degree in Bachelor of Computer Science from New York University, US, the M.Sc. degree in Bachelor of Science (Computer), Universiti Kebangsaan Malaysia, and the Doctor of Philosophy in Information Technology from Tun Hussein Universiti Onn Malaysia (UTHM), Malaysia. He is the Principal Investigator, Faculty of Computer Science and Information Technology, Center of Information Security Research (CISR), Universiti Tun Hussein Onn Malaysia (UTHM). His areas of expertise are information, computer and communications technology (ICT), security systems, and security services (including digital forensics, steganography, network security, public key infrastructure, and biometrics). His research area of interest was watermarking, and his innovation fields of interest are GPS-based applications and the internet of things (IoT). He can be contacted at email: shamsulk@uthm.edu.my.









**Dr. Ezak Fadzrin Bin Ahmad Shaubari**     his higher studies in MS Food Engineering by coursework at the at the department of Food Science, University of Leeds from 1981-1982. He is attached to the Faculty of Biotechnology and Biomolecular Sciences. His research area then was on spray drying of food. With a small research grant provided by UPM, he developed the process for producing spry-dried coconut milk which made the national headlines. His vast experience and expertise in the field of biotechnology and biomolecular sciences have enabled him to become a national point of reference in the area of biomass, renewable energy and waste utilization. He has also served as a consultant to the Science Advisor Office, Prime Minister's Department, on the national project on biomass utilisation and is the national representative for the Asia Biomass Association headquartered in Tokyo, Japan. He can be contacted at email: ezak@uthm.edu.my.



**Sukri, S.T., M.Kom.**     is currently lecturers at Abdurrah University, Indonesia. She obtained her Master in Information Technology from the Indonesian Islamic University, Indonesia. He is currently pursuing his Ph.D. in Doctor of Philosophy of Information Technology at Tun Hussein Universiti Onn Malaysia (UTHM). His research interests include data mining and machine learning. He can be contacted at email: sukri@univrab.ac.id.



**Zul Indra, M.Sc.**     is a lecturer at the University of Riau, Indonesia. He obtained his Master in Computer and Information Sciences from Universiti Teknologi PETRONAS. He is currently pursuing his Ph.D. in computer science. His research interests include software development, data mining, and machine learning. He can be contacted at email: zulindra@lecturer.unri.ac.id.