Text Mining Approach for Topic Modeling of Corpus Al Qur'an in Indonesian Translation

1st Dwi Rolliawati[1], 2nd Indri Sudanawati Rozas[1], 3rd Khalid[1], 4thIndri Sudanawati Rozas[1]
{dwi_roll@uinsby.ac.id[1], indrisrozas@uinsby.ac.id[1], khalid@uinsby.ac.id[1], mratodi@uinsby.ac.id[1]}

UIN Sunan Ampel, Faculty of Science and Technology, Surabaya, Indonesia[1]

Abstract. Qur'an is a religious text for Moslem that is revealed to humanity as a guide to solve any problems in all aspects of life. Therefore Quranic text is widely translated in various countries around the world, including in Indonesia which is predominantly inhabited by Moslem. Difficulties in understanding the Arabic Quranic text as well as still limited research on the Indonesian translated Quran in accordance to science and technology, have opened a broad challenge to contribute to this realm. This paper proposed topic modelling of corpus in Indonesian Translated Quran by generating four main topics that were closely/firmly related to human life: 1) heaven (surga) and hell (neraka), 2) World (dunia) and Afterlife (akhirat), 3) Science (ilmu), charity (amal) and jihad, 4) Day (siang), night (malam), life (hidup), and death (mati). The research method was divided into two stages, the text mining stage and the topic modeling stage. The moderator variables for this research were defined as Makki and Madani, as terms referring to the revelation location of Quranic verses. Of all the 4 modeling topics tested by word count, Makki'sSurahs contribute above 50% compared to Madani'sSurahs. In conclusion, the study results hopefully can be benefited as a convincing contribution from science's point of view that Makki's verses are indeed emphasizing the faith as the foundation of Islam. This can be seen from number frequencies of the words "hidup" (161), "neraka" (157), "surga" (105), "dunia" (127), "amal" which are closely related to human faith in life were mentioned, discussed and elaborated more in Makki's verses than in Madani's.

Keywords: component; Indonesian Translation Quran, Makki, Madani, topic modeling, corpus

1 Introduction

The Quran is a significant religious text written in Quranic Arabic which is followed and used as guidance by the Islamic faith believers. The Quran in the sense of language means "perfect reading" that is revealed to the human as a guide to solve any problems in all aspects of life. Indonesia as the largest Muslim country in the world has great potential in grounding the Quran thus the Quran can be more easily learned, understood and practiced. The Indonesian Quran translation has been circulated widely in the broader community both in print and digital. However, the Indonesian translation of the Quran still has not represented the convenience in searching of particular topic required for a specific purpose. Meanings and ideas are overlapping from ayah to ayah and from surah to surah, therefore drawing out an implied connections would need more in-depth study and time for discovering the hidden thematic structures [1]. As Quran's function as a guide for humans, research on the Qur'an continues to grow in various fields. Research by Zakariah et al. have studied the future trends,

ICONQUHAS 2018, October 02-04, Bandung, Indonesia

DOI 10.4108/eai.2-10-2018.2295559

review, and analysis for the development of research on Quran, ranging from security for digital Quran, e-learning and the implementation of Natural Language Processing (NLP), etc. [2]. Topic modeling is a hot field of study in both machine learning and NLP. Topic models are generative models that based on probability distributions of multiple topics in a document over a set of words [3].

This research present topic modeling of corpus Quran in Indonesian translation. We considered Makki and Madanisurahs as the variable for topic modeling categorization. The reason for taking the subject theme of Makki and Madani in this research is because there is still limited study examine such subject, and the implementation of Makki and Madani are merely only for Surah's categorization and Surah's content verification. Details of the previous researches are described in section 2. Hopefully, this study could contribute well to the science community and for future research as well.

2 State of The Art

2.1 Literature Review

Text Mining

The most common text mining approach involves a representation of text that based on keywords. A keyword-based methodology can be combined with other statistical elements (machine learning and pattern recognition techniques, for example) to discover relationships between different aspects in the text by recognizing repetitive patterns which are presented in the content. Text Mining is also known as Intelligent Text Analysis or Knowledge Discovery in Text or Text Data Mining which incorporates areas such as information retrieval, information extraction, data mining, computational linguistics and natural language processing [4]. In general, the text processing in text mining has:

(i) Text pre-processing, in this stage we conducted tokenization, stop-word removal and stemming process.
(ii) Text transformation, in this stage text document represented by the words it contains and their occurrences. Two approaches used for document representation are (i) bag of words (ii) vector spaces.
(iii) Feature Selection, which eliminates redundant and irrelevant features.
(iv) Clustering and Classification text

Quranic Resource

The Quran was revealed through the Prophet Muhammad (PBUH) within 22 years two months and 22 days [5] in the city of Mecca and Medina. The Mecca revelation period lasted for 12 years of Prophet Muhammad's prophetic period and the revealed Surahs that descend on this time belong to Makkah's Chapter, whilst the Madinah revelation period which was began since the Hijrah event, lasted for ten years and all the revealed Surahs during this period is called Madani's Chapter. The original Arabic Quran has been distributed in online digital form for some time, but only few books elaborate vocabulary complexity of the Quran. The Quran contained 6,236 verses (ayahs) and divided into 114 chapters. Each of ayahs is in form of a string of words. The total Quranic text comprises of about 77,000 words in the Arabic language [1]. Whilst the complete Indonesian translation Quranic text before stemming are ranges from 151,236 words to 37,857 words after derivation.

Topic Modeling

Topic models are computer algorithms that identify potential patterns of word occurrence using the distribution of words in a collection of documents [6]. Topic modeling is an unsupervised learning method based on the idea that a large group of records may accurately be classified into a small number of topics. Topic models generate interpretably, semantically coherent issues, which can be examined by enumerating the most likely words for each subject [7]. The output of topic modeling is a set of topics consisting of clusters of words that co-occur in these documents according to specific patterns Topic models are useful for a variety of tasks such as organization, classification, collaborative filtering and information retrieval [8]. There are several approaches to implement topic modeling. The following are some of the most popular topic modeling related to approaches addressed in information retrieval and machine learning literature [9]:

(i) Latent Semantic Indexing (LSI)
(ii) The unigrams mixture model,
(iii) Probabilistic Latent Semantic Indexing (PLSI)
(iv) Latent Dirichlet Allocation (LDA)

2.2 Previous Research

Many research on text mining and topic modeling applied to Quran have been conducted during these recent two years, and apparently would still going to be trending topic up to years ahead. Text mining implementation on Quran by Sharaf in his research has produced a semantic annotation of the Quran by using N-Gram from the original language of the Qur'an [10]. This research also created two files of the raw Qur'an, one which contains only makki chapters, and another only madani chapters. And the result showed that Makki's chapters contains 47,643 words (61.2%) of them 6,358 hapax legomena (13.3%) whilst Madani's chapters contains 30,161 words (38.8%) of them 4,621 hapax legomena (15.3%). Another research study that had used the Quran Indonesian translation was merely a literature review on the SQA System applications of text mining to ITQ (Indonesian Quran Translation) [11]. Alhawarat et al. in his research on the Al-Quran text processing implementation, when it was viewed from Most Frequent Words which is visualized in form of a word cloud using TF-IDF weighting, still used the original language without going through the stemming process [12]. As related to the Quran modeling topic, research by Panju had also used TF-IDF and factorization in topics visualization [13], while Alhawarat in his research has used the LDA method (The Latent Dirichlet Allocation) to extract the modeling topic based on the Yusuf's Surah datasets only [3]. Still, on the LDA method, a similar study was carried out by Siddiqui et al. that described 15 most frequent terms from Quran after normalization and corpus specific stopwords removal[1]. Through probabilistic methods, the Surahs were classified and mapped into Makki or Madani and came to the conclusion that MakkiSurahs emphasized more on the basic tenets of the religion including oneness of Allah, the prophet-hood of Muhammad PBUH and the coming of Judgment Day. Whilst the Madanisurahs laid down on Islamic law and jurisprudence, outlining ritualistic aspects, moral and ethical codes, laws of governance, etc. The above-mentioned previous researches, mostly still used the original Arabic text as its corpus for the text mining management. As for the use of Indonesia, Quran translation was limited only to the word searching [14][15] and the utilization of its

architecture for question answering (QA) system development on relevant documents of Indonesian Quran translation[16].

3 Research Methods

In general, the research method was divided into two stages, the text mining stage and the topic modeling stage. The final result of the text mining stage was word-count (which was later being modeled with word cloud), whilst the final outcome of topic modeling stage was the visualization of topics intended to model. Overall, the two steps of the study were summarized below in Figure 1:

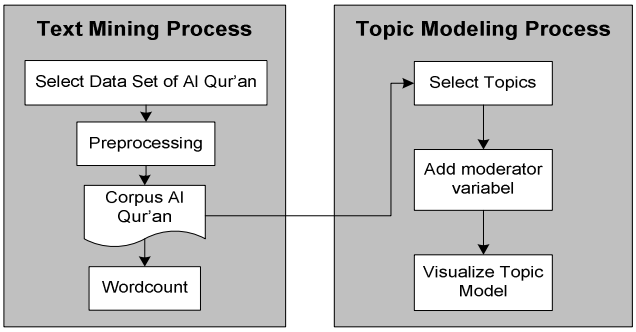

Figure 1. Research Method

3.1 Text Mining

As mentioned earlier, we use the dataset obtained from the qurandatabase.org site in the form of *.csv file extension. Before the Quran corpus was processed and visualized, it was necessary to conduct a preprocessing to obtain a clean dataset. One of the steps in preprocessing was data clearance from stop-word. Stop-word list considered as common words that did not provide any vital information, such as the words "yang," "di," "kemudian", and so forth. The stop-word used in this research came from a literary stop word [17]. Until this phase, the corpus results can resume at the topic modeling stage.

To view the text mining processing result from the Quran corpus content can be seen in the following table (Table 1).

Table 1. Corpus in Indonesian Quran Translation

Data Table

Data instances: 42953
Features: 1
Meta attributes: 1

	Word	Word Count
1	allah	3394.000
2	maha	1018.000
3	beriman	652.000
4	tuhan	614.000
5	rasul	478.000
6	bumi	462.000
7	kafir	444.000
8	ayat	437.000
9	manusia	436.000
10	azab	376.000
11	allah maha	368.000
12	tuhanmu	323.000
13	langit	321.000
14	kitab	310.000
15	quran	303.000
16	jalan	296.000
17	tanda	263.000
18	petunjuk	251.000
19	neraka	249.000
20	kaum	245.000
21	hamba	241.000
22	laki	235.000
23	berbuat	228.000
24	musa	227.000
25	ayat ayat	210.000
26	malaikat	206.000
27	tiada	193.000
28	zalim	193.000
29	hati	191.000
30	surga	188.000
31	langit bumi	187.000
32	dunia	181.000
33	dosa	178.000
34	engkau	176.000
35	allah allah	167.000
36	sisi	165.000
37	tuhanku	165.000
38	kiamat	162.000
39	nyata	160.000
40	peringatan	154.000
41	menciptakan	154.000
42	rahmat	154.000
43	air	149.000
44	mendustakan	148.000
45	syaitan	142.000
46	negeri	141.000
47	harta	141.000
48	perempuan	139.000
49	golongan	138.000
50	takut	135.000
51	muhammad	135.000
52	perbuatan	135.000
53	bertakwa	134.000
54	tuhannya	132.000
55	kepadaku	132.000
56	kebenaran	131.000
57	nikmat	130.000
58	akhirat	127.000
59	penyayang	126.000
60	nabi	126.000
61	muka	125.000
62	saleh	123.000
63	umat	121.000
64	mendengar	120.000
65	agama	120.000
66	laki laki	120.000
67	maha penyayang	119.000
68	mati	118.000

From the table above, when the modeling applied is in the word cloud format, the result is as shown in Figure 2 below:

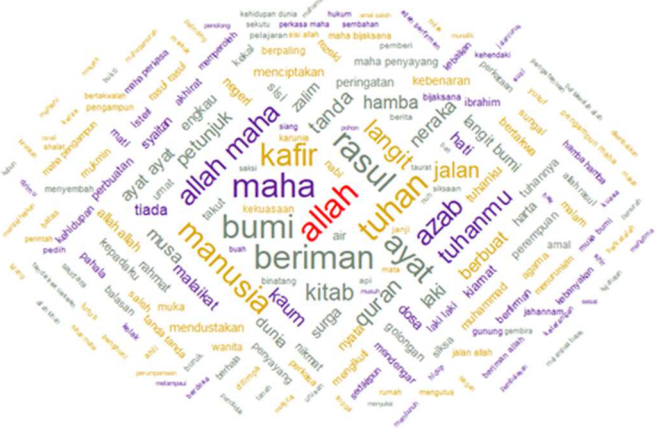

Figure 2. Topic modeling word cloud

3.2 Topic Modeling Process

The topic modeling process was started with selecting the topics to be modeled, then selecting the moderator variable and finally calculating and visualizing the chosen model. There was four topic modeling conducted in this research, which were as follow:

(i) Heaven (surga) and hell (neraka)
(ii) World and hereafter (duniadanakhirah)
(iii) Science, charity, and jihad (ilmu, amal, jihad)
(iv) Day, night, life, and death (siang, malam, hidup dan mati)

The above four topics used similar moderator variable, both Makki and Madani verses. The range of verses as the object of this research study is defined as the first verse up to verse 6236 of the Qur'an. The visualization result and the elaboration of four types of topic modeling were discussed in the following section.

4 Result

4.1 Topic modeling on heaven and hell

We intended to do the modeling into verses in Quran containingwords "surga" and "neraka". In Figure 3 we tried visualized the first verse up to verse 6236 of the Quran, where the red dots indicate verses containing the word "neraka," and the blue dots indicate verses containing the word "surga." The results showed that the "surga" and "neraka" topicwere more likely discussed in Makki verses compared to Madani verses.

In detail, the word "surga" in the Quranic corpus mentioned 160 times and 223 times for the word "neraka." As for the proportion,the word "surga" was mentioned 105 times in Makki verses and 55 times in Madani verses, while the word "neraka" was mentioned 157 times in Makki verses and 66 times in Madani verses.

Such topic modeling generated from this research shall interesting to be elaborated further particularly in relation to the verses origin “why and wherefores” (*asbabunnuzul*), as to examines the fact that Makki verses explain more the “surga” and “neraka” topic than Madani verses.

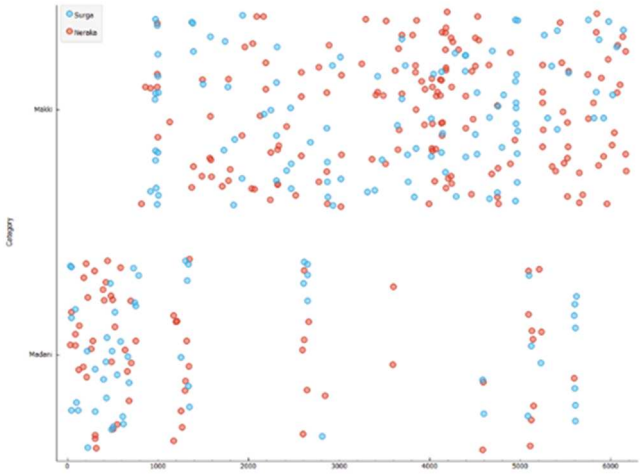

Figure 3. Distribution of heaven and hell topic

4.2 Topic modeling on the world and the hereafter

On this topic modeling, we found that the word "dunia" was 182 times mentioned, and the word "akhirat" was 54 times mentioned. As shown in Fig. 4, the red dots indicate verses containing the word "dunia," and the blue dots indicate verses containing the word "akhirah." At further examination with moderator variables in both verses, turning out that the word "dunia" 55 times mentioned in Madani'sverses while in Makki's verses was 127 times mentioned. Then for the word "akhirah" was mentioned as much as 54 times in the Quran corpus, with the distribution as much as 40 times mentioned in Makki verses and 14 times in Madani. Again, further investigation on *asbabunnuzul* regarding this topic modeling is necessarily substantial.

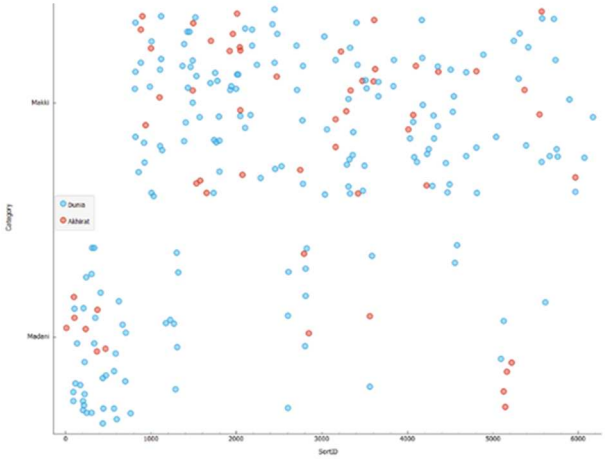

Figure 4. Distribution of world and hereafter

4.3 Topic modeling on science, charity, and *jihad*

We were making attempt to count and visualize verses in Quran that contain the words "ilmu," "amal", and "*jihad*." From the entire verses examined the word "ilmu" was 50 times appeared, with 34 of them were found in Makki's verses, and the other 16 founded on Madani's verses. In searching of the word "amal," we found that the word has appeared 141 times, of which 92 of them were found in Makki's verse, while the other 49 were in Madani'sverses. As for the word "*Jihad*," the word was found in 31 verses, with 27 of them were included in Madani's verses and the rest of 4 belongs to Makki's verses. Figure 5 below is a visualization of topic modeling on "ilmu", "amal", and *jihad*, where the red dots indicates Makki's verses and the blue dots shows the verses of Madani

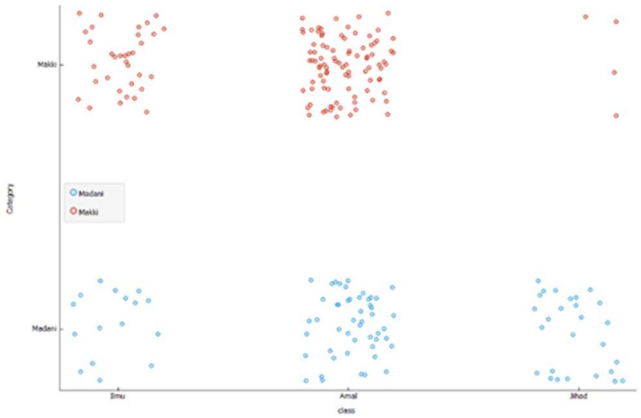

Figure 5. Distribution of science, charity, and *jihad*

4.4 Topic modeling on the day, night, life, and death

In this topic, we found the word "siang" as much as 53 words, 45 on Makki verses and the others 8 were in Madani verses. For the word "malam," as many as 51 words successfully detected with 45 of them found in Makki verses and 6 on the Madani's verses. As for the word "hidup," we have successfully detected 204 words on the Quranic corpus, with 161 of them were found in Makki verses, while the rest 43 words were found in Madani verses. And for the last, for the word "mati" we have detected 107 words in the Quranic corpus, with 61 of them were found in Makki's verses, while the 46 others were found in Madani verses. It appeared that the terms "siang" and "malam" were relatively balanced, but the words "hidup" and "mati" mentioned in ratio 2 to 1.

Figure 6 below is a visualization of topic modeling on "siang," "malam," "hidup," and "mati," where the red dots indicate Makki's verses and the blue dots shows the verses of Madani

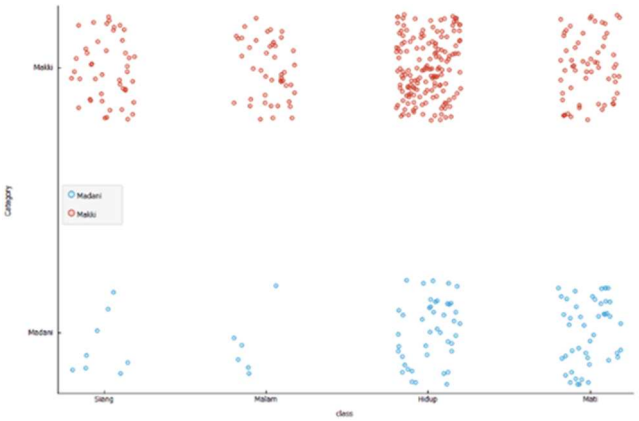

Figure 6. Distribution of day, night, life, and death

Based the four main topics that had been modeled, we can see that the words "hidup", "neraka", "surga", "dunia", and "amal" that are closely related to human faith in life were much more elaborated and discussed (see Table II). The analysis results also had shown thatMakki's verse were more dominating in all the words tested (see Table III).

Table 2. Topic Modeling Result Based on Word Count

Variabel	Topic 1		Topic 2		Topic 3			Topic 4			
	surga	neraka	dunia	akhirat	ilmu	amal	jihad	siang	malam	hidup	mati
Makki	105	157	127	40	34	92	4	45	45	161	61
Madani	55	66	55	14	16	49	27	8	6	43	46
total	160	223	182	54	50	141	31	53	51	204	107

Table 3. Topic Modeling Result Percentages View

Variabel	Topic 1		Topic 2		Topic 3			Topic 4			
	surga	neraka	dunia	akhirat	ilmu	amal	jihad	siang	malam	hidup	mati
Makki	65,63%	70,40%	69,78%	74,07%	68,00%	65,25%	12,90%	84,91%	88,24%	78,92%	57,01%
Madani	34,38%	29,60%	30,22%	25,93%	32,00%	34,75%	87,10%	15,09%	11,76%	21,08%	42,99%

These research findings prove empirically that Makki's verse give more emphasis to the topic of faith, whereas in the Madani's verses the words "hell", "heaven", "world", "charity" and "life" respectively have been discussed more than the other words. However, the result was less empirically able to prove that the verse Madani's did give emphasis on the law, *muamalah* and so forth. A further study need to be conducted to develop and obtain more scientific proof.

References

[1] M. A. Siddiqui, S. M. Faraz, and S. A. Sattar, "Discovering the Thematic Structure of the Quran using Probabilistic Topic Model," *Proc. - 2013 Taibah Univ. Int. Conf. Adv. Inf. Technol. Holy Quran Its Sci. NOORIC 2013*, no. December, pp. 234–239, 2015.

[2] M. Zakariah, M. K. Khan, O. Tayan, and K. Salah, "Digital Quran Computing: Review, Classification, and Trend Analysis," *Arab. J. Sci. Eng.*, pp. 1–26, 2017.

[3] M. Alhawarat, "Extracting Topics from the Holy Quran Using Generative Models," *Int. J. Adv. Comput. Sci. Appl. - See more http//thesai.org/Publications/ViewPaper?Volume=6&Issue=12&Code=ijacsa&SerialNo=38#sthash.7kmJYsB9.dpuf*, vol. 6, no. 12, 2015.

[4] M. Sumathy, K.L.; Chidambaram, "Text Mining: Concepts, Applications, Tools and Issues - An Overview," *Int. J. Comput. Appl.*, vol. 80, no. 4, pp. 29–32, 2013.

[5] M. Shihab, "Membumikan Al-Quran," *Bandung: Mizan*, no. November, pp. 1–232, 1992.

[6] C. Jacobi, W. Van Atteveldt, and K. Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling," *Digit. Journal.*, vol. 4, no. 1, pp. 89–106, 2016.

[7] Z. Cao, "A Novel Neural Topic Model and Its Supervised Extension," *Proc. Twenty-Ninth AAAI Conf. Artif. Intell. A*, pp. 2210–2216, 2015.

[8] J. W. Uys, N. D. Du Preez, and E. W. Uys, "Leveraging unstructured information using topic modelling," *PICMET Portl. Int. Cent. Manag. Eng. Technol. Proc.*, no. c, pp. 955–961, 2008.

[9] A. Zinman *et al.*, "Latent dirichlet allocation," *MIS Q.*, vol. 3, no. 3, pp. 993–1022, 2010.

[10] A. M. Sharaf, "The Qur'an Annotation for Text Mining," *Rev. Lit. Arts Am.*, no. December, 2009.

[11] S. J. Putra, T. Mantoro, and M. N. Gunawan, "Text mining for Indonesian translation of the Quran: A systematic review," *2017 Int. Conf. Comput. Eng. Des.*, pp. 1–5, 2017.

[12] M. Alhawarat, M. Hegazi, and A. Hilal, "Processing the Text of the Holy Quran: a Text Mining Study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 2, pp. 2–7, 2015.

[13] M. H. Panju, "Statistical Extraction and Visualization of Topics in the Qur'an Corpus," *Student.Math.Uwaterloo.Ca*, 2014.

[14] D. B. Nugraheni, M. A. Bijaksana, and E. Darmawiyanto, "Analisis Dan Implementasi Pencarian Kata Berbasis Konkordansi Dan N-Gram Pada Terjemahan Al-Quran Berbahasa Indonesia Analysis And Implementation Concordance Search And N-Gram For Words In Al-Quran English Translation," vol. 4, no. 3, pp. 4713–4718, 2017.

[15] A. Herdianto, "Pencarian Ayat-Ayat Alquran Berdasarkan Konten Menggunakan Text Mining Berbasis Aplikasi Desktop," *Pencarian Ayat-Ayat Alquran Berdasarkan Konten Menggunakan Text Min. Berbas. Apl. Deskt.*, vol. 2, pp. 1–14, 2010.

[16] 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," *2014 Int. Conf. Cyber IT Serv. Manag. CITSM 2014*, pp. 104–107, 2014.

[17] J. Demšar *et al.*, "Orange: Data Mining Toolbox in Python," *J. Mach. Learn. Res.*, vol. 14, p. 23492353, 2013.