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

1st Dwi Rolliawati
Faculty of Science and Technology
UIN Sunan Ampel
Surabaya, Indonesia
dwi_roll@uinsby.ac.id

2nd Indri Sudanawati Rozas
Faculty of Science and Technology
UIN Sunan Ampel
Surabaya, Indonesia
indrisrozas@uinsby.ac.id

3rd Khalid
Faculty of Science and Technology
UIN Sunan Ampel
Surabaya, Indonesia
khalid@uinsby.ac.id

4th Muhamad Ratodi
Faculty of Science and Technology
UIN Sunan Ampel
Surabaya, Indonesia
mratodi@uinsby.ac.id

Abstract- *The Qur'an is the religious text for Muslims 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 predominantly by Muslim. Difficulties in understanding the Quranic text in Arabic as well as the limited research on the Indonesian translation Quran related to science and technology, have opened a broad challenge to contribute to this realm. This paper proposed topic modelling of corpus in Indonesian Translation Quran by generated four main topics that are firmly related to human life, such as 1) heaven (surga) and hell (neraka), 2) The world (dunia) and the hereafter (akhirat), 3) Science (ilmu), charity (amal), and jihad, 4) Day (siang), night (malam), life (hidup), and death (mati). Those four topics were related to the moderator variables associated with the revelation location of Quranic verses (Makki and Madani). Of all the modeling topics tested by word count, Makki's Surahs contributes above 50% compared to Madani's Surahs. So the study results can be a reinforcement from the science's point of view that Makki verses were indeed emphasizing the faith as the foundation of Islam. This can be seen from the frequencies numbers that indicate the words "hidup" (161), "neraka" (157), "surga" (105), "dunia" (127), "amal" which is closely related to the human faith during their life in the world was discussed more in Makki's verses than Madani's.*

Keywords-component; *Indonesian Translation Quran , Makki, Madani, topic modelling, corpus*

I. INTRODUCTION

The Quran is a significant religious text written in Quranic Arabic and followed by the Islamic faith believers. The Quran in the sense of language means "perfect reading" that is revealed to 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 so that 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 ease of searching in a particular topic required for a specific purpose. Meanings and ideas overlap from ayah to ayah and from surah to surah and therefore drawing out the 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, 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 Madani surahs as the variable for topic modeling categorization. Why Makki and Madani themes? This because there was not much research that exposes more about it as well as Makki and Madani implementation only comes to the Surah categorization and the Surah content verification only. Details of the previous research described in section 2. For that reasons, we hope this study can contribute to the science community and future research.

II. STATE OF THE ART

A. 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 is present 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:

1. Text pre-processing, in this stage we conducted tokenization, stop word removal and stemming process.
2. 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.
3. Feature Selection, which eliminates redundant and irrelevant features.
4. Clustering and Classification text

Quranic Resources

The Quran was revealed through the Prophet Muhammad in 22 years two months 22 days [5] in the city of Mecca and Medina. The period of Mecca lasted for 12 years of Prophet Muhammad's prophetic period and the letters that descend on this time belong to Makkah's Chapter, while the period of Madinah which began since the Hijrah event lasted for ten years and the Surah that revealed on this period is called Madani Chapter. The original Arabic Quran has been circulating in digital form online for some time, but there were a few books that only explain difficult vocabulary of the Quran. The Quran contained 6,236 numbered verses (ayahs) and divided into 114 chapters. Each of ayahs is a string of words. The total Quranic text comprises of about 77,000 words in the Arabic language [1]. While the complete Quranic text translation of Indonesia before stemming ranges from 151,236 words and about 37,857 words after deriving.

Topic Modelling

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 classify into a small number of topics. Topic models generate interpretable, semantically coherent issues, which can examine 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 approaches addressed in information retrieval and machine learning literature [9]:

1. Latent Semantic Indexing (LSI)
2. The unigrams mixture model,
3. Probabilistic Latent Semantic Indexing (PLSI)
4. Latent Dirichlet Allocation (LDA)

B. Previous Research

Research on text mining and topic modeling with the Quran as its subject has been going on in the last two years and will still be trending topic until several years ahead. Implementation of text mining on al-Quran by Sharaf in his research has produced a semantic annotation of the Quran using N-Gram from the original language of the Qur'an [10]. From the research also generated Makki chapters contained 47,643 words (61.2%) of them 6,358 hapax legomena (13.3%) and Madani chapters 30,161 words (38.8%) of them 4,621 hapax legomena (15.3%). Another study that has used the Indonesian Quran translation was merely a literature review on the SQA System applications of text mining to ITQ (Indonesian Quran Translation) [11]. Alhawarat et al. research on the implementation of Al-Quran text processing viewed from Most Frequent Words have used TF-IDF weighting to visualize it in the form of a word cloud, although still used the original language without going through the stemming process [12]. Related to the Quran modeling topic, research by Panju has also used TF-IDF and factorization in topics visualization [13], while Alhawarat in his research has used the LDA method (The Latent Dirichlet Allocation) in extracting 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 describes 15 most frequent terms from the Quran after normalization and corpus specific stopwords removal[1]. Through probabilistic methods, the Surah included in Makki and Madani verses mapped as a form of proof that Makki surahs emphasized the basic tenets of the religion including oneness of Allah, the prophet-hood of Muhammad PBUH and the coming of the Day of Judgment. The Madani surahs laid down Islamic law and jurisprudence, outlining ritualistic aspects, moral and ethical codes, laws of governance, etc. From some of the previous researches presented, most of the research uses 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 search implementation [14][15] and offers architecture for question answering (QA) system development on relevant documents of Indonesian Quran translation[16].

.

III. RESEARCH METHODS

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

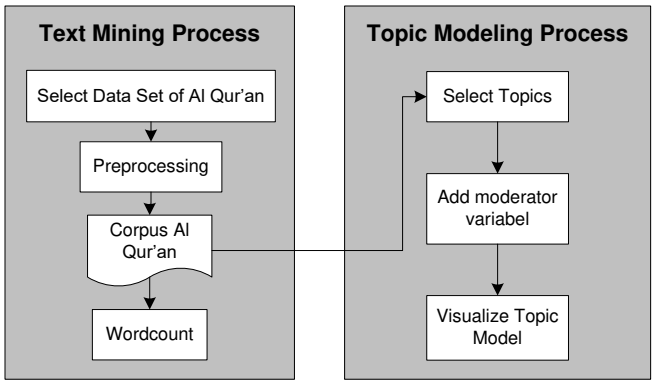

Fig. 1. Research Method

a. Text Mining

As mentioned earlier, we used the dataset obtained from the qurandatabase.org site in the form of *.csv file extension. Before the Quran corpus processed and visualized, it was necessary to conduct a preprocessing to get a clean dataset. One of the steps in preprocessing was data clearing 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 comes from a literary stop word [17]. Until this phase, the corpus results can resume at the topic modeling stage.

To view the text mining process 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

If from the table above we make the modeling in the word cloud format, the result as shown in Figure 2 below:

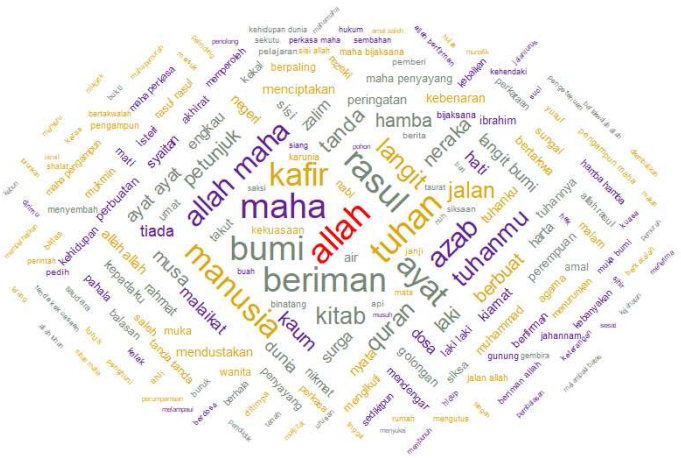

Fig. 2. Topic modeling word cloud

b. Topic Modeling Process

The topic modeling process begins 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 is related to:
1. Heaven (surga) and hell (neraka)
2. The world and the hereafter (dunia dan akhirat)
3. Science, charity, and jihad (ilmu, amal, jihad)
4. Day, night, life, and death (siang, malam, hidup dan mati)

The four topics use the same moderator variable, both Makki and Madani verses. The range of verses that are the object of our study is the first verse up to verse 6236 of the Qur'an. The four process of topics visualization result detail discussed in the following section.

IV. RESULT

1. Topic modeling on heaven and hell

We were interested in modeling verses that contain the words "surga" and "neraka" in the Qur'an. 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.

This generated topic modeling become interesting to be discussed further related to the verses origin (*asbabun nuzul*), which examines the fact that Makki verses explain the topic more than Madani verses.

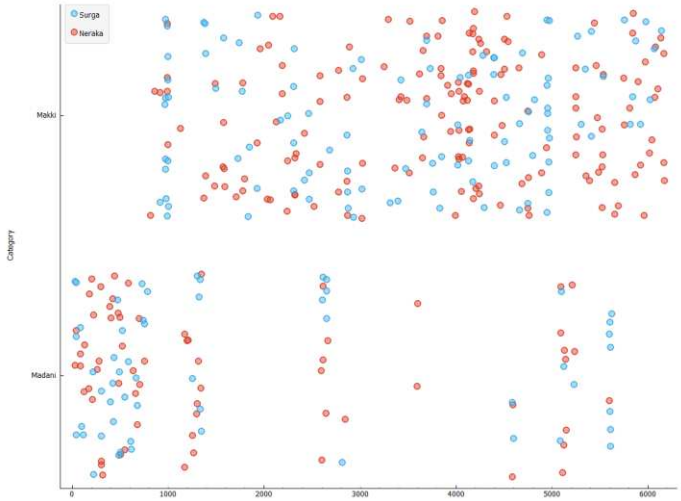

Fig. 3. Distribution of heaven and hell topic

2. Topic modeling on the world and the hereafter.

On this topic modeling, we have found that the word "dunia" was 182 times mentioned, and the word "akhirat" was 54 times mentioned. As showed in Fig. 4, the red dots indicate verses containing the word "dunia," and the blue dots indicate verses containing the word "akhirat." At further examination with moderator variables in both verses, turned out that the word "dunia" 55 times mentioned in Madani verses while in Makki verses was 127 times mentioned. Then for the word "akhirat" 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 asbabun nuzul regarding this topic modeling is needed.

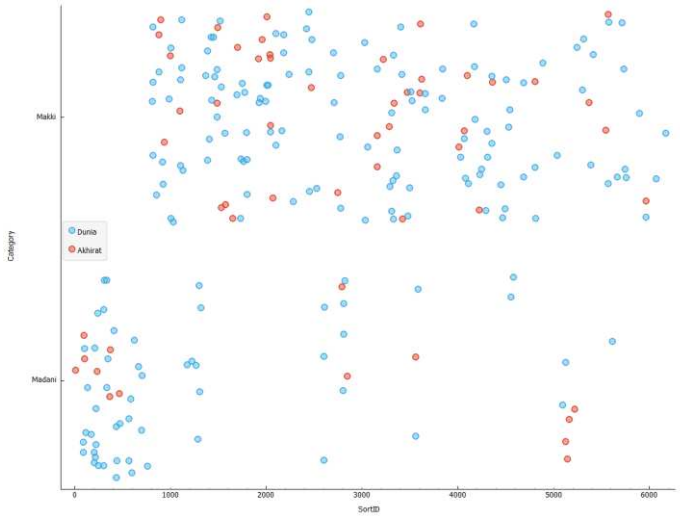

Fig. 4. Distribution of word and the hereafter

3. Topic modeling on science, charity, and *jihad*

We tried to count and visualize verses that contain the words "ilmu," "amal", and "*jihad*." From the entire verses examined, the word "ilmu" was 50 times appeared, with 34 of them have been found in Makki verses, and the other 16 founded on Madani verses. In search of the word "amal," we have 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 Medina verses. 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

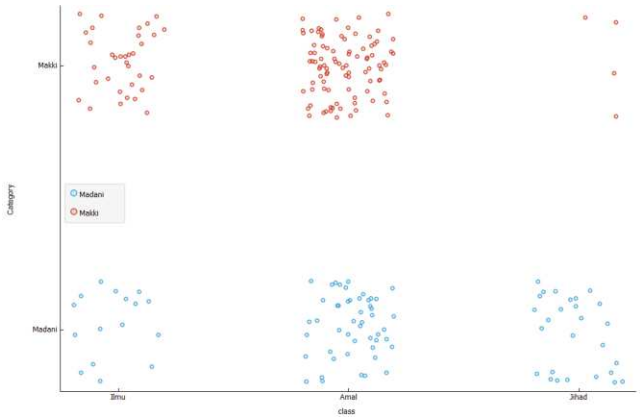

Fig. 5. Distribution of science, charity, and *jihad*

4. Topic modeling on the day, night, life, and death.

On this topic, we found the word "siang" as much as 53 words, 45 on Makki verses and eight others on Madani verses. For the word "malam," as many as 51 words successfully detected with 45 of them found in Makki verses and six on the Madani 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) found in Madani verses. And the last, for the word "mati" we have detected 107 words on the Quranic corpus, with 61 of them were found in Makki verses, while the 46 others found in Madani verses. It appears that the terms "siang" and "malam" were relatively balanced, but the words "hidup" and "mati" mentioned with 2 to 1 ratio. In 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

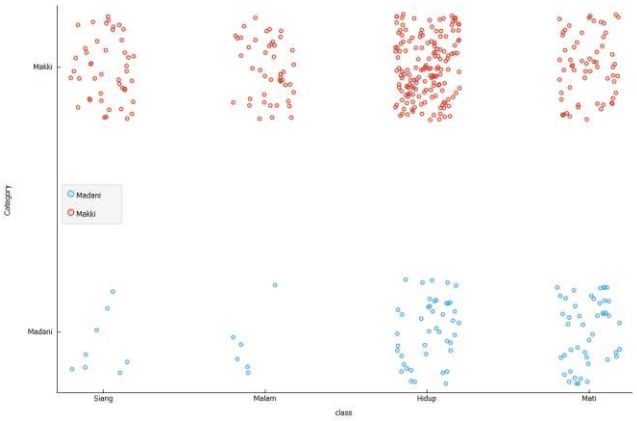

Fig. 6. Distribution of day, night, life, and death

Based the four main topics that have been modeled, we can see that the words "hidup", "neraka", "surga", "dunia", "amal" that closely related to human faith during their life in the world more likely to be discussed (see table 2). The analysis results also have shown Makki's verse dominating on all the words tested (see table 3).

Table 2 Topic Modelling 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	27	45	45	161	61
Madani	55	66	55	14	16	49	4	8	6	43	46
total	160	223	182	54	50	141	31	53	51	204	107

Table 3 Topic Modelling Result in percentage's 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%	87,10%	84,91%	88,24%	78,92%	57,01%
Madani	34,38%	29,60%	30,22%	25,93%	32,00%	34,75%	12,90%	15,09%	11,76%	21,08%	42,99%

This findings proves empirically that Makki's verse does give more emphasis to the topic of faith, whereas in the Madani's verses the words "hell", "heaven", "world", "charity" and "life" sequently 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. Furthermore, we will try to develop research for more perfect 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.