

# Semantic Model Evaluation Dataset For Indonesian In Al-Qur'an Vocabulary: Similarity And Relatedness

Falia Amalia  
Program Studi S1 Informatika  
Universitas Telkom  
Bandung, Jawa Barat, Indonesia  
faliaamalia@student.telkomuniversity.ac.id

Moch Arif Bijaksana  
Program Studi S1 Informatika  
Universitas Telkom  
Bandung, Jawa Barat, Indonesia  
arifbijaksana@telkomuniversity.ac.id

**Abstract** — In the Qur'an, a lot of words can be used to be researched especially in the case of Natural Language Processing (NLP), one of which is the Model semantic distribution for common words and linkages as in this research. Previous research, which uses the Qur'an dataset only performed semantic similarities to certain surah, evaluation of datasets similarities and semantic linkages but the application of the Turkish language, and application of evaluation of datasets for verbs only. The purpose of the study was to provide insight into the researcher's semantic model by evaluating the model in several attributes. Balance the pairs of words of the Quranic vocabulary dataset for a class of nouns and verbs with human frequencies to evaluate the durability of a semantic model regarding the problem of rare word occurrence. The word pair is selected to enable evaluation of a semantic model distribution with multiple Word attributes and link pairs of words such as the frequency of word occurrence, concreteness, and relationship type (e.g., synonymy, antonymy). The Dataset consists of 500 pairs of words and is assisted by 15 respondents, of which each pair has two distinct values for similarity and relevance. The method used is the Sim-Rel vector, questionnaire, and the calculation of gold standard, until the result of performance calculations using a correlation of Spearman Rank of 0.909. The Sim-Rel vector axis Gets the result with 4 areas i.e. SU = 23 pairs of words, SR = 77, DU = 192, and DR = 208.

**Keywords** — Natural Language Processing; Distribution Semantic Model; Sim-Rel Vector; Spearman Rank

**Abstrak** — Dalam Al-Qur'an, sangat banyak kata yang dapat digunakan untuk diteliti khususnya dalam hal *Natural Language Processing* (NLP), salah satunya yaitu Model Semantik Distribusi untuk kesamaan kata dan keterkaitan seperti pada penelitian ini. Penelitian sebelumnya, yang menggunakan dataset Al-Qur'an hanya dilakukan kesamaan semantik semantik untuk surah tertentu, evaluasi dataset kesamaan dan keterkaitan semantik namun penerapan pada bahasa Turki, dan penerapan evaluasi dataset untuk kata kerja saja. Tujuan dari penelitian ini adalah untuk memberikan wawasan kepada peneliti model semantik dengan mengevaluasi model dalam beberapa atribut. Menyeimbangkan pasang kata dataset kosa kata Al-Qur'an untuk kelas kata benda dan kata kerja dengan frekuensi manusia untuk mengevaluasi kekokohan model semantik mengenai masalah kemunculan kata langka. Pasangan kata dipilih untuk mengaktifkan evaluasi distribusi model semantik dengan beberapa atribut kata dan hubungan pasang kata seperti frekuensi kemunculan kata, konkret dan jenis hubungan (misalnya, synonymy,

antonymy). Dataset ini terdiri dari 500 pasang kata dan dibantu oleh 15 responden, yang setiap pasangan memiliki dua nilai yang berbeda untuk kesamaan dan keterkaitan. Metode yang digunakan adalah Sim-Rel vektor, kuesioner dan perhitungan *gold standard*, hingga hasil perhitungan performansi menggunakan korelasi Spearman Rank sebesar 0,909. Sumbu vektor Sim-Rel mendapatkan hasil dengan 4 daerah yaitu SU = 23 pasang kata, SR = 77, DU = 192, and DR = 208.

**Keywords** — *Natural Language Processing*; Model Semantik Distribusi; *Sim-Rel Vector*; *Spearman Rank*

## INTRODUCTION

Semantic relatedness and similarities relate to the linguistic field, especially in Natural Language Processing (NLP) [1], which has become an interesting and widely researched topic. Semantic and word-related similarities have important roles in some of the tasks of NLP and some associated fields such as text classification, document clustering, text summarization, and so forth [2]. Semantic similarities are a process used to estimate the strength of a semantic relationship between a unit of language, a concept, or a common sense of the word. Evaluate similarities in documents widely used, one of them for application related to information search, natural language processing, etc. [3]. Evaluation of Word similarity (i.e., WordSim) is one of the oldest natural methods of assessment of the semantic distribution model.

The Qur'an is a holy book for Muslims and is the most prominent guideline and source of law. The Qur'an has 30 Juz, 114 surahs, and 6236 verses [2]. Quran is the holy book for Muslims who believe and follow the religion of Islam. This book is currently being used as the guidebook of life to 1.6 billion Muslims in the world living today[4]. In 6236 these verses many vocabularies have similarities and interrelated [5], but many of them are far apart the range is not even found in one surah or one juz, therefore it is necessary to study deeper to understand the meaning associated with the word Qur'an. Therefore, it appears the idea of the evaluation of datasets of value in common and the relation of semantic words to the pairs of Qur'anic words such as between the words "Allah" and "Tuhan"; "Surga" and "Neraka" or even "Nabi" and "Rasul" by going through numerical descriptions obtained by the comparison of meaningful supporting information or describing nature.

In this paper, will present works in the similarity of semantic words by utilizing computational linguistic learning, especially in the field of semantic similarities and the distribution of semantic models. Distribution Semantic (DS) is a theory related to linguistic tradition, which according to [6] in the proposal of Zellig Harris, from the analysis of distribution as a fundamental linguistic rock as a scientific discipline. The linguistic elements in two (orthogonal) types are a lot of learning with a distribution relationship in Semantic research distribution (DS) today: syntagmatic and paradigmatic. In research [7], in Milton's opinion, to measure the mastered vocabulary cannot be obtained when the word is written only as a lexical word, but can be measured when spread in sentences. One example is synonyms such as "nabi" and "rasul" in the sentence "Muhammad sebagai [nabi / rasul] Allah ", where two words tend to happen in the same sentence. In previous research using the Qur'an, the dataset is only performed semantics for certain surah [8], the relation of words with Arabic vocabulary [9]. Other studies, discussing the development of datasets similarities and semantic relatedness but for application in the Turkish language [10], and the application of evaluation of datasets for verbs only [11]. According to [12] there are other studies whose data is still in use today such as SimLex-999, WordSim353, SimLex-666. The SimLex-999 is still being used to date as one of the gold standards in semantic distribution (DSM) modeling research.

To complement the previous research, this study makes a list of the words in the Qur'an with the form of a noun class and a verb that presents the similarity of Word and word relation (i.e., association). Evaluation of this dataset aims to provide a field of semantic modeling for the Indonesian language derived from the vocabulary of the Qur'an with an intrinsic evaluation source based on the definition of the word obtained from the *Kamus Besar Bahasa Indonesia* (KBBI). Seeing the development of the existing semantic sciences to study the Qur'an, this research was created at the same time to create datasets, so that future research with the focus of other discussions can use this dataset to help the research progress. Like the research [10], [11], the level of similarity of this word will be measured in the continuous range [0.10] for each pre-selected pair of words. A score of 10 shows similarities and maximum relatedness, while 0 does not show similarities and relatedness. The value of similarity and association is obtained by involving the gold standard that the results are described in the vector functions and relationship type functions. The built-in evaluation is expected to produce a good performance based on the calculated correlation value. The correlation value in question is derived from the calculation of the Spearman Rank. The challenge of this research is to mimic human intuition in measuring the similarity and interconnectedness of words in semantics.

#### DESIGN MOTIVATION

This section, explains the difference in Word evaluation on previous research, explaining some common word evaluation issues and design decisions made to address those issues in research, and calculations for the emergence of words that according to the study were scarce seen from some parameters. Contributions from this paper include knowing more about the difference between the similarities and

relatedness of the word evaluation dataset for Indonesia Languages, the main objective is to balance each pair of words by several morphological and semantic attributes, analysis and visualization of words similarity and relatedness of datasets that have value for each word pair, and finish it by reviewing the gold standard and the correlation value that exists.

#### A. The Difference in Word Evaluation

After the theoretical differences of paradigmatic and syntagmatic relationship types, one can easily apply those differences to the word evaluation by assuming that "similarity is a paradigm and relatedness that is a type of relationship that is syntagmatic.". In the study [11] thoroughly explained the difference between similarity and association (i.e., Sim-Rel) caused the problem of the dataset's limitations. The SimVerb-3500 [11] also defines criteria for the evaluation dataset in three: representative, clearly-defined, consistent, and reliable. According to [10], most WordSim datasets such as RG65, WordSim-353, and MEN do not meet clearly defined criteria because screen guidelines use "similarity" and "linkage", and the word "association" is in one another.

Semantic modeling for the Indonesian language with intrinsic evaluation resources targeting the encouraged morphological problems caused by the rich agglutinative properties of languages. Since our research aims to collect both similarity and linkage scores from participants, we provide clear-defined detailed instructions on the questionnaire screen (Figure 2-4).

#### B. The Emergence of Rare Words

In this evaluation, the emergence of rare words that appear in the translation of the word Qur'an is quite a lot. Since the dataset [10], [11], [13] measure a rare word level, the study received a 26% yield on [10] for the number of word occurrences of the range 0-320. The rare value of this study is assumed by a total of 2913 words, taken the middle value of the data to 1456 to the end, it shows the number of rare word occurrence only under 3.

### METHODOLOGY

#### A. Sources Dataset

The Dataset used in this final task comes from the online website e-book. A dataset containing a translation of the Qur'an by 6236 verses or lines that want to know the value of similarities and relatedness of pairs of words taken for the noun class and verbs in the Indonesian Language.

#### B. Steps of Research

The stages of this research are:

1. The system converts the whole verse translation into a word token that will be in the next process
2. The system selects candidates for identical words according to the class
3. Manually, create 500 pairs of words which already include nouns and verbs
4. Questionnaire and calculating the gold standard
5. The system will render the data plot according to the value obtained

- The system will calculate performance values with the correlation of Spearman Rank

### C. Sim-Rel Vector Space

As is the reference paper, this paper will try to use the method on the paper to collect the pairs of words in the search for the value of similarity, and the association is the Sim-Rel Vector Space. This paper will try to prove whether the Indonesian vocabulary derived from the Qur'an using the formula of the function, later can be used or not.

The x-axis represents the association with the r score, and the y-axis represents similarity to the s score of each word pair in each dataset. Group division is marked with, SU (similar-not related); SR (similar-related); DU (different-not related); DR (different-related) is a categorical label from a possible sub-space semantic or *ss*, then  $ss = f_1(r, s)$  function would be,

$$ss = f_1(r, s) = \begin{cases} SU, & \text{if } s \geq 5 \text{ and } r < 5 \\ SR, & \text{if } s \geq 5 \text{ and } r \geq 5 \\ DU, & \text{if } s < 5 \text{ and } r < 5 \\ DR, & \text{if } s < 5 \text{ and } r \geq 5 \end{cases} \quad (1)$$

Then the method also provides a function formula with  $t = 2$  indicates a variable threshold that represents the boundary point relation type of space where synonyms, antonym, irrelevant is a categorical label possible semantic relation-types *rt*,  $rt = f_2(r, s)$  function will,

$$f_2(r, s) = \begin{cases} synonym, & \text{if } 10 - t \leq s \text{ and } 10 - t \leq r \\ antonym, & \text{if } 10 - t \leq r \text{ and } s \leq t \\ irrelevant, & \text{if } t \geq r \text{ and } t \geq s \end{cases} \quad (2)$$

Word-pairs could be accepted as synonyms if their *rt* value is assigned to synonym varying by the *t* parameter. The same rule applies to the irrelevant value, too. Intuitively select a threshold of  $t = 2$  values for Sim-Rel semantic space. The  $t = 2$  parameter equals all axes and relationships with the model and the simplicity of visualization.

### D. Gold Standard

Gold Standard is a result of the personal opinion that is a reference in the process of measuring semantic similarities between a text pair and a word on an absolute scale. The predominant gold standards for semantic evaluation in NLP do not measure the ability of models to reflect similarity [13]. The personal advice here is obtained from 15 respondents who have contributed to filling the questionnaire.

### E. Correlations Spearman Rank

Spearman rank correlation coefficient  $r_s$  is a nonparametric or distribution-free rank statistical measure of the strength and the direction of the arbitrary monotonic association between two ranked variables or one ranked variable and one measurement variable [14]. Formula  $r_s$  for the correlation of Spearman are as follows:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (3)$$

When  $d_i = rg(A_i) - rg(B_i)$  is the significant difference between the two ranked variables and  $n$  is the total number of observations [15].

## RESULT AND DISCUSSION

### A. Preprocessing

Translation of the Qur'an in Bahasa Indonesia in the process to be a word token that omitted its word attributes such as punctuation, equalization letters, cutting out unnecessary words because it prevents many of the emergences of unused words. As a result of this stage, the system gets a word count of 61683 words with all the word classes read by the name label.

### B. Word Candidate Selection

A total of 6236 verses or about 80000 words in the translation are chosen based on the type of class they say are nouns and verbs. The system calculates the overall number of PostTagging for the noun and verb from the beginning of 36285 with 24794 for nouns and 11491 for verbs. Then the system calculates to determine the identical word for each class of the word, and the result is 4731 with an identical noun as many as 2904 words, while 1827 for the identical verb. Then, manually last selection to match the word according to KBBI and generate 2193 for nouns and 1733 for a verb.

In [10] research for word candidates, the study resulted in 639 words for nouns only derived from the Turkish vocabulary mix and vocabulary of other datasets, while the research was pure from the Indonesian translation dataset of the Qur'an.

Statistics Chart Design words for the dataset can be seen in Figure 1, and below are examples of candidates taken based on the number of words from the largest, which was also previously matched using KBBI.

Table 1. Top 10 Noun Candidate With The Greatest Number Of Values

ID	Identical Word Candidates			
	Arab	English	Indonesian	Count
18	الله	allah	allah	2878
17	إله	lord	tuhan	969
22	ذَا	people	orang	672
5	الْيَوْمِ	day	hari	483
1458	إِلَّا	except	kecuali	433
27	الْأَرْضِ	earth	bumi	416
21	عِقَاب	punishment	hukuman	364
63	الْخَوْفِ	fear	ketakutan	282
1055	خَيْرًا	good	kebaikan	259
72	الْحَقِّ	truth	kebenaran	210

Table 2. Top 10 Verb Candidate With The Greatest Number Of Values

ID	Identical Word Candidates			
	Arab	English	Indonesian	Count
10	عَدَّة	said	bilang	843
139	تَقُولُوا	say	mengatakan	756
6	تُؤْمِنُ	believe	percaya	422
266	عَلِيم	know	tahu	279
146	أَتَى	come	datang	245
176	بَصَرَ	see	lihat	235
425	يُعْطِي	give	memberi	193
330	يُضْرِبُ	Make	membuat	192
9	تستخدم	used	digunakan	191
311	يُحِبُّ	Like	menyukai	186

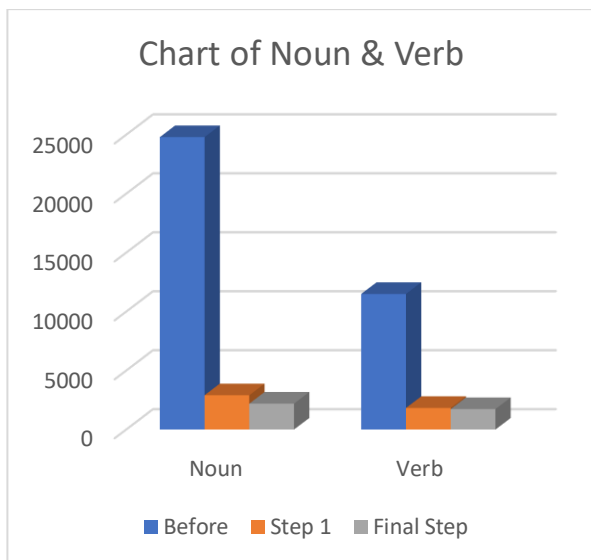


Figure 1 Chart Design Statistics For Noun & Verb

C. Pairing-Word Selection

From the identical word obtained, the system calculates the number of occurrences of each word with the translated document. For similar research, the final SimVerb-3500 data set contains 3.500 verb pairs in total, covering all associated verb [11]. Similarly, the research [10] generates 500 pairs of nouns.

Due to time and budget constraints, this study sets the target size of the project dataset to 1.000 scores (500 word-pairs) on nouns and verbs. To be a comparison, most of the WordSim datasets in previous research have fewer word pairs, (SimLex-999 = 999, RG = 65, WordSim-353 = 353, RW = 2034, MEN = 3000) [10].

Therefore, the word already has a sorted value of the word that has the most number of benefits and takes the top 250, and this applies to verbs and nouns. Then, randomly selected to pair from any word obtained with the noun and verb rules,

not mix. Below are seven samples of manually created word pairs.

Table 3. 5 Examples of Word Pair Results

Word Pair Results	
Word Pair (ENG)	Word Pair (IND)
Adam – humankind	Adam – manusia
Guidance – Quran	Petunjuk – Quran
Allah – god	Allah – tuhan
Book – Quran	Buku – Quran
Punishment – penalty	Hukuman – siksa
Prophets – messenger	Nabi – rasul
Friend – relatives	Teman – kerabat

This research plans the process of constructing datasets in three phases: Word candidate, KBBI, and Word pair selection.

Table 4. Three Stages of Dataset Construction

	Stage 1	Stage 2	Stage 3
	Word Candidates	KBBI	Word Pair
Goals	Using existing data	To customize word candidates with KBBI	Randomly balancing pairs of words
Input	Indonesia Translation Collection	The output from Stage 1	The output from Stage 2
Process	Tag Word type for each word and KBBI	Match word one by one into KBBI online	Mapping pairs (every word used 2-5 times building word-pairs)
Output	2904 (N) + 1827 (V)	2193 (N) + 1733 (V)	500 (N&V)

If viewed, the table above only contains two stages, because all the vocabulary used is taken from the translation of the Quran that appears, hence the absence of the word outside the source of the Qur'an used. After all, the Qur'an is a definite and irreversible database, other than that the goal is to create a dataset from the vocabulary of the Qur'an in Indonesia Language.

D. Questionnaire

Looking at the [10], [11], [13] the study also uses questionnaires to get the value of the human versatile by adjusting the conditions and situations. These conditions and circumstances that cause data retrieval differ from other studies.

1. Media Questionnaire

This research uses Spreadsheets or Excel online, which can be accessed by respondents who do not know the time. The reasons for using this include:

1. Situations and conditions require everyone to work from home.



2. Preventing the delay of charging from the specified duration
3. Prevent the respondent from discharging the questionnaire
4. The charging of respondents can always be controlled by the authors.

Here's an image that demonstrates the design of Excel, created to collect data through human annotators using Spreadsheets :

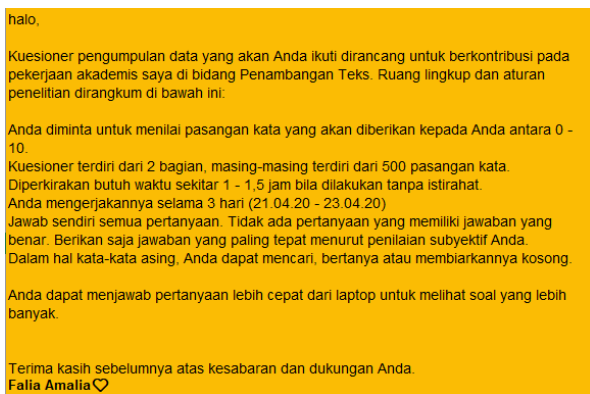


Figure 2 Greetings Of Questionnaire

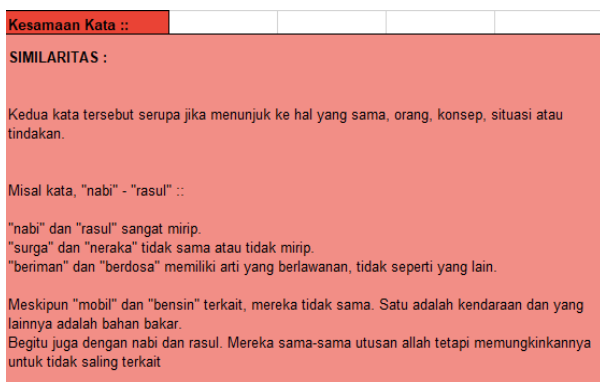


Figure 3 Rules of the Word similarity Questionnaire

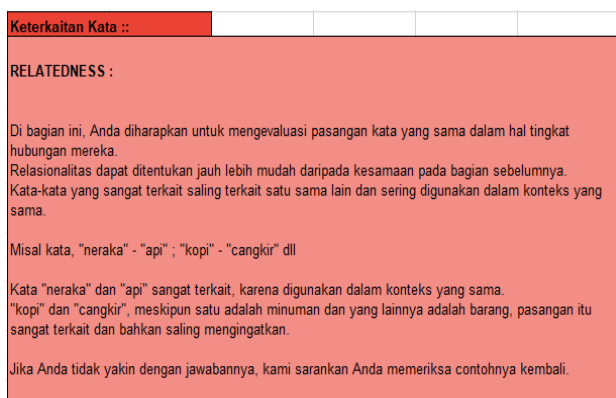


Figure 4 Rules of the Word Relatedness Questionnaire

NO	Pasang Kata	Pasang Kata IND	Nama Lengkap : Talitha Farah Kurniawati Kampus : Telkom University Semester : Tingkat Akhir		Nama Lengkap : Dean Arya P. Kampus : Universitas Trisakti Semester : 8	
			Similaritas atau Kesamaan	Relatedness atau Keterkaitan	S	R
1	Adam - mankind	Adam - manusia	4	4	7	6
2	guidance - Quran	petunjuk - Alquran	7	4	9	7
3	Allah - God	allah - tuhan	10	2	10	10
4	book - Quran	buku - Alquran	9	5	4	3
5	punishment - penalty	hukuman - siksa	8	3	7	7
6	prophets - Messenger	nabi - rasul	9	2	6	5
7	friend - relatives	teman - kerabat	9	2	7	5
8	guide - favors	orang yg mendapat petunjuk - nikmat	6	6	6	6
9	fear - victory	ketakutan - kemenangan	4	7	2	8
10	mercy - curse	rahmat - laknat	7	0	0	0

Figure 5 Questionnaire From Respondents

## 2. Respondents

The respondent's problem has more influence on this questionnaire. Due to the situation and conditions described above, this study took data from 15 respondents who came from students with different majors and universities. By accident, 12 are female respondents, and the rest are males, with 12 being in the final semester and the remainder under semester 5. A sample image filling questionnaire from respondents can be seen in Figure 5.

## E. Gold Standard Calculation

This Gold Standard calculation seeks the value of all the benefits of similarities and relatedness obtained from 15 respondents, so it has two benefits of Gold Standard. Then, this a gold standard value is inserted into the function formula (1). Below are the sample results of the gold standard for 13 pairs of Indonesian words.

Table 5. 13 Examples of Gold Standard Calculation

Word1	Word2	GS.S	GS.R
adam	manusia	18.66666667	19
Petunjuk	quran	21.66666667	21.66666667
allah	tuhan	24.5	22.66666667
buku	quran	14.33333333	14.66666667
Hukuman	Siksa	16.33333333	16.5
nabi	rasul	18.16666667	17.83333333
teman	kerabat	19.33333333	18.16666667
Petunjuk	nikmat	12.66666667	16
Ketakutan	kemenangan	7.66666667	11.16666667
Rahmat	Laknat	5.83333333	4.66666667
Setan	Jin	16.5	16.33333333
Adam	Rasul	17.5	16.33333333
Jibril	manusia	3.33333333	14.83333333

GS. S (Sim) is the gold standard for similarity value, during GS. R (Rel) is the gold standard for relatedness value.

## F. Dataset Analysis

Analysis of the obtained data shows the Sim-Rel vector with the assumption and configuration of the respondent. Apparently, from the Sim-Rel function, get good results also if tried in a state of the Indonesia translation of the Qur'an. Deploying plots based on the r and s values of gold standard produces similar visuals. Here's the result:

1. Sub-space SU get the result as many as 23 pairs of words; SR = 77; DU = 192; and DR = 208.

- From the results based on the human annotator, it generates the relationship type for Synonym = 16; Antonym = 11 and Irrelevant = 473
- If viewed from a plot spread, many of the dots overlap or have the same coordinate point of 5.93 for the GR and 1.33 for GS, such as "Jibril" – "Manusia" and "berbicara" – "melihat" (see table 6)
- Below is a plot distribution for each type of relationship that the capture system represents the other word pairs

In research [10] get the value of the SU = 0 which means that if researched, a pair of words that have a similar word should the word can't relate to each other because it has a high value of similarities, such as "wanita" and "perempuan". These two words have a high value of similarity and should not be related. Meanwhile, in this study vector, SU = 23-word pairs. This is the difference that can be compared due to respondents in this study with previously different studies, allowing for different results as well. Similarly, for values in other vector axes of the study, with SR = 52, DU = 215, and DR = 233.

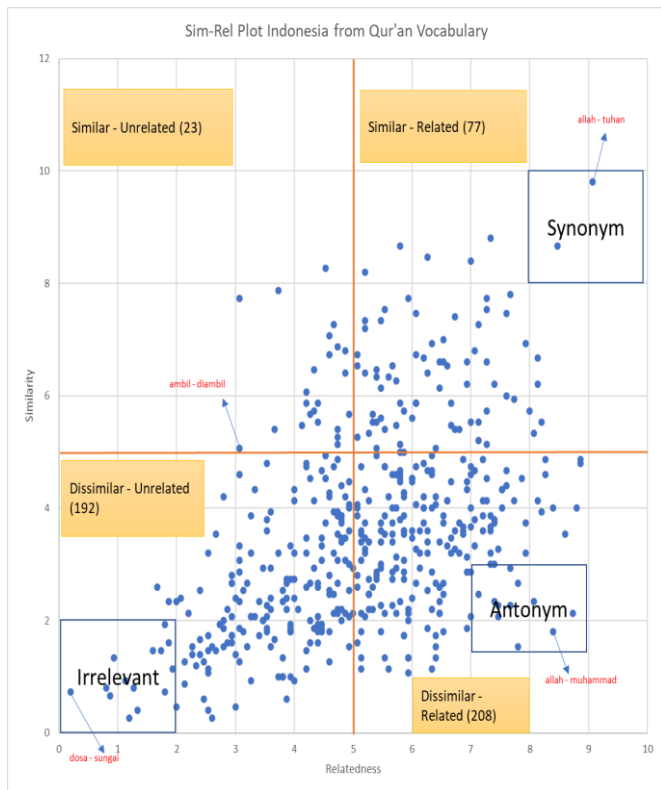


Figure 6. Scatter Plot Indonesian

Table 6. 9 Sample Word-Pairs From The Final Dataset.

Word1	Word2	Sim	Rel	SS	Relation Type
allah	tuhan	9.07	9.8	SR	Synonym
Petunjuk	quran	21.67	21.17	SR	Synonym
Rahmat	Laknat	7.67	11.17	SU	Synonym
allah	muhammad	8.4	1.8	DR	Antonym
ambil	diambil	3.06	5.06	SU	Irrelevant

jibril	manusia	5.93	1.33	DR	Synonym
berbicara	melihat	5.93	1.33	DR	Irrelevant
kebajikan	kelemahan	3.93	1.8	DU	Antonym
dosa	sungai	0.2	0.73	DU	Irrelevant

Sample word-pairs from the final dataset (Sim = value of similarity, Rel = value of relatedness, SS = semantic sub-space)

### G. Correlation Spearman Rank

In this study, for the final stage, calculating the performance value using the correlation of Spearman Rank. The system calculates the correlation value for all values obtained between similarities and relatedness and gets a result of 0.909, which is still better than the most word similarity datasets SimLex-999 = 0.67 [13] AnlamVer = 0.748, (WS-Sim=0.667, MEN=0.68) [10] or SimVerb-3500 = 0.628 [11].

### CONCLUSION

This research represents a dataset for the evaluation of the semantic model for Indonesian Languages derived from the Qur'an vocabulary for the class of nouns and verbs with the type of relation. Reviewing the main objectives of this research is to balance pairs of words from human frequency datasets to evaluate the durability of a semantic model on the issue of rare word occurrence. In the discussion about the appearance of the word rare above, the noun is obtained for the number of occurrences of rare words under three as many as 739 words or about 25.36%. For verbs, the number of occurrences of the word is less than equal to three 867 words or approximately 50.02% according to the parameters. The number of rare word occurrences of a noun is not much different than the result of an AnlamVer dataset. Using a vector Sim-Rel method, this method succeeded in evaluating the semantic model for Indonesian Languages in the vocabulary of the Qur'an to get better results compared to research in Turkish, although some results are not as expected because it involves the value of respondents. This is evidenced by the greater correlation value gained even by the research of other evaluation datasets and with a slightly more number of respondents.

Hopefully, the dataset about the similarity and relationship of this word can be useful to support other research in the future with a different discussion so that more research discussing the vocabulary Qur'an. From this research anyway, for future research, it can develop larger datasets. However, when conducting the development or development of datasets, it is advisable that people who are experts in their field such as experts in a particular language or Qur'an to be able to reproduce other types of relations according to different morphology of Indonesia, for example, hypernym, hyponym, and the meronym, etc.

### AWARD

Apart from the Lord, parents, and lecturers, thank you so much for the 15 respondents who are willing to contribute in judging the pairs of words that have been made according to what they understand, understand their simplicity, and support the continuity of the research on this paper.

REFERENCES

- [1] F. T. Admojo and E. Winarko, "Sistem Pencarian Informasi Berbasis Ontologi untuk Jalur Pendakian Gunung Menggunakan Query Bahasa Alami dengan Penyajian Peta Interaktif," *IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 10, no. 1, p. 23, 2017, doi: 10.22146/ijccs.11186.*
- [2] F. F. Guntara, M. A. Bijaksana, and A. F. Huda, "Pembangunan Daftar Kata Terkait pada Kosa Kata Al-Qur'an Berdasarkan Kesamaan Distribusional," *J. Tek. Inform. dan Sist. Inf., vol. 6, no. 2, pp. 139–146, 2020, doi: https://doi.org/10.35957/jatisi.v6i2.192.*
- [3] E. Supriyati and M. Iqbal, "Pengukuran Similarity Tema Pada Juz 30 Al Qur'an Menggunakan Teks Klasifikasi," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput., vol. 9, no. 1, pp. 361–370, 2018, doi: 10.24176/simet.v9i1.1955.*
- [4] D. A. Wiranata, M. A. Bijaksana, and M. S. Mubarak, "Quranic concepts similarity based on lexical database," in *2018 6th International Conference on Information and Communication Technology, ICoICT 2018, 2018, vol. 0, no. c, pp. 264–268, doi: 10.1109/ICoICT.2018.8528794.*
- [5] G. R. Prabowo, "Pembangunan Monolingual Word Alignment Pada Terjemahan Al-Quran Berbahasa Indonesia," *J. Linguist. Komputasional, vol. 1, no. 2, p. 51, 2018, doi: 10.26418/jlk.v1i2.11.*
- [6] G. Benotto, E. Giovannetti, and O. Nahli, "An application of distributional semantics for the analysis of the Holy Quran," *Colloq. Inf. Sci. Technol. Cist, pp. 374–379, 2017, doi: 10.1109/CIST.2016.7805074.*
- [7] Y. Nirmalasari, A. S. Ghazali, and G. Susanto, "Pemerolehan Kata Pemelajar BIPA UM Ditinjau dari Segi Sintagmatik dan Paradigmatik," *J. Pendidik. Teor. Penelitian, dan Pengemb., vol. 3, no. 1, pp. 82–99, 2018.*
- [8] N. M. Abdelaal and S. M. Rashid, "Semantic loss in the holy Qur'an translation with special reference to surah Al-WaqiAAa (Chapter of the event inevitable)," *SAGE Open, vol. 5, no. 4, 2015, doi: 10.1177/2158244015605880.*
- [9] H. Hamam, M. T. Ben Othman, A. Kilani, F. Ncibi, and M. Ben Ammar, "Exploring Qur'an by using Aspects and Dependencies," *J. Adv. Comput. Sci. Technol. Res., vol. 5, no. 4, pp. 133–141, 2015.*
- [10] G. Ercan and O. T. Yildiz, "AnlamVer: Semantic Model Evaluation Dataset for Turkish - Word Similarity and Relatedness," in *Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 3819–3836, [Online]. Available: http://www.gokhanercan.com/anlamver%0Ahttps://www.aclweb.org/anthology/C18-1323.*
- [11] D. Gerz, I. Vulic, F. Hill, R. Reichart, and A. Korhonen, "Simverb-3500: A large-scale evaluation set of verb similarity," in *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings, 2016, pp. 2173–2182.*
- [12] T. Kliegr and O. Zamazal, "Antonyms are similar: Towards paradigmatic association approach to rating similarity in SimLex-999 and WordSim-353," *Data Knowl. Eng., vol. 115, pp. 174–193, 2018, doi: 10.1016/j.datak.2018.03.004.*
- [13] F. Hill, R. Reichart, and A. Korhonen, "SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation," *Computational Linguistics, vol. 41, no. 4, pp. 665–695, 2015, doi: 10.1162/COLI.*
- [14] C. Xiao, J. Ye, R. M. Esteves, and C. Rong, "Using Spearman's correlation coefficients for exploratory data analysis on big dataset," *Concurr. Comput. Pract. Exp., vol. 22, no. 6, pp. 685–701, 2015, doi: 10.1002/cpe.*
- [15] A. K. J and S. Abirami, "Aspect-based opinion ranking framework for product reviews using a Spearman's rank correlation coefficient method," *Inf. Sci. (Ny), vol. 460–461, pp. 23–41, 2018, doi: 10.1016/j.ins.2018.05.003.*