Vol. 7, Issue 2, pp. 119-128 | DECEMBER 2022
DOI: https://doi.org/10.53840/myjict7-2-164

MEASURING SIMILARITY ENGLISH TRANSLATIONS OF SURAH AL-FATIHAH USING COMPUTATIONAL METHODS

[i*]Kholed Langsari

Faculty of Science and Technology Fatoni University

*(Corresponding author) e-mail: langsari@ftu.ac.th

Article history:
Submission date: 12 October 2022
Received in revised form: 24 November 2022
Acceptance date: 28 November 2022
Available online: 31 December 2022

Keywords:
Al-Fatihah, Text Mining, TF-IDF, Cosine similarity

Funding:
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interest:
The author(s) have declared that no competing interests exist.

Cite as:
Kholed Langsari. (2022). Measuring Similarity English Translations of Surah Al-Fatihah using Computational Methods. *Malaysian Journal of Information and Communication Technology (MyJICT), 7*(2), 119-128.
https://doi.org/10.53840/myjict7-2-164

Abstract

The similarity measurement between different English translations of Surah Al-Fatihah using four computational methods is performed. This study was conducted using Bag of Words, Term-Frequency (TF), Term-Frequency–Inverse Document Frequency (TF-IDF) and Cosine Similarity Latent Semantic Indexing (LSI). In the process, I used four translators of Surah Al Fatihah (ibn Kathir, Yusuf Ali, al Saadi, Al Tabari). This paper proposed an approach to discover similarity links between four translations of Surah Al-Fatihah, using the cosine similarity proposed method. The results from the proposed method serve new valuable in similarity measuring between the English translations of Surah Al Fatihah. The presented results could be used for classification of English translations and comparison for future translations and understanding of the Holy Quran word.

Introduction

The Qur'an is the holy book of Islam. It is composed of Allah's exposure to the prophet Muhammad, transmitted through the angel Gabriel and it is the source of Islamic religious practice and devotion. The Qur'an is written down in Arabic that consists of 114 chapters, each chapter is known as Surah and each chapter consists of Verses known as an ayah (Matt, 2019). According to the Qur'an, there are 86 surahs were given to Muhammad the prophet of Islam while he was in Mecca mostly related to the social relation and growth of the Islamic community. The other 28 surahs were revealed in Medina concerned with faith and the afterlife (Mohammed, 2005).
Surah Al-Fatihah (in Arabic: الفاتحة, meaning "The Opening" or "The Opener") is the first chapter or surah of the Quran. It consists of seven verses that are a prayer for the guidance, path, lordship and mercy of Allah. This chapter has a crucial position in Islamic prayer (salat). The name Al-Fatihah ("the Opener or the opening") is the exact subject matter of the surah. Fatihah is used which open a subject or a book or a ceremony and any other in Islamic culture. In other words, is a sort of introduction. The word الفاتحة originally from the root word فتح which means to open, explain, disclose, keys of treasure, etc. That means Surah Al-Fatihah is the beginning and summary of the whole holy Quran (Jannah, 2019).

However, The holy Qur'an was primarily written in the Arabic language. And the Qur'an has been translated into many others languages for those who do not speak or understand Arabic. Salman the Persian was the first translator of the Qur'an, who translated the Surah al-Fatihah into the Persian language during the early 7th century. Currently, there have translated more than 114 languages. Moreover English translations there have been numerous translators. Translating the Qur'an is not an effortless one; many native Arab speakers will confirm that some Quranic passages are fatiguing to understand even in the original Arabic script. A part of this is the natural difficulty of any translation; in Arabic, as in other languages, a single word can have a variety of words to translate such as the word الرَّحمن in English can be translated to the word *"merciful"*, *"gracious"* or *"beneficent"*

In this study, I use a computational linguistic approach to assess and measure the similarity of the English translation of the Qur'an. In the arithmetic approach, the English translations are evaluated using Cosine Similarity. I evaluate the similarity of four translators based on three computational methods from the natural language processing (NLP) domain; there Preprocessing pipeline, Term Frequency-Inverse Document Frequency or TF-IDF and Cosine Similarity. The results determine whether one English translation is how much similar and links with another English translation based on the given criteria (Matt, 2019).
This paper is organized as follows; Section 2 presents the related works by other researchers in the same scope. Section 3 presents the methodology of this paper to discover the similarity between translations. Section 4 presents the experiment result of the proposed approach using a case study. And Section 5 gives a conclusion and remarks on future works.

Related Work

Many researchers have reviewed the Holy Quran English translations, including (Kidwai, 1987) (Nassimi, 2005) and (Muhammad, 2008). First Kidwai reviewed and suggested several English translations of the Qur'an. Secondly, Muhammad provided some critical views of both English translations, but some translations were for a certain group of people such as the Shia (Husain, 1831) or Ahmadiyya (Robinson, 1997). There were many factors such as ideological bias, Arabic fluency (Al- Azzawi and Z. H. Alwan, 2013), English proficiency, and knowledge of Islamic law, because of these many factors, each translation is different and unique. Third, Nassimi made an objective comparison of English translations of the holy Qur'an.

These studies have shown similarities and different perspectives between English translations. In this study, I use a mathematical computational method called similarity scale to find the similarity between English translations of the Qur'an the similarity scale (Jurafsky and Martin, 2008) contains many applications such as information retrieval, natural language processing and query analysis. For instance, we can use a similarity measurement scale to compare two or more documents to evaluate if the documents are the same. Another example is in the query analysis configuration. When the user queries "supercar", the method can suggest and recommend a similar query such as "high-performance vehicle" or "expensive car" (Jurafsky and Martin, 2008). The similarity measurement in this paper is based on similarity measures from the domain the computational linguistics (CL)or natural language processing (NLP). There are many similarities in measures metrics for comparing documents, sentences, and words.

There is a method called the bag of words (bows) model. This model is based on the set theory concept idea from the mathematics discipline. In the bows model, the set carries unordered words with no duplications of each other. The similarity measuring is a ratio value between the distance and the union between the two sets. There is no required external information used in the bows model. Another method is the Term-Frequency or TF model. This model projects the terms and their time of appearance (frequencies) into the vectors model in a vector space. Then, the similarity evaluation is calculated using the Cosine Distance of the vectors. Another method is the Term-Frequency Inverse Document Frequency (TF-IDF) model. Different from the two previous models, this model uses a corpus as a dictionary vocabulary. The basic concept idea is that a corpus will provide a better representation vocabulary of the terms used in the model. This model used the term-document (TD) matrix to acquire the weight for each term. These weights will be projected into a vector space as cosine similarity. Then, the similarity measure is calculated using the Cosine Distance of the vectors. The latent semantic indexing (LSI) model uses the concept idea that there are latent or hidden relationships between the terms in the documents (Deerwester et al.,1990) (Dumais et al., 1988). This latent similarity can be computed using a Single Value Decomposition (SVD) method. This model also uses a vocabularycorpus for matching. The LSI model extends the TF-IDF model by using different terms and weights calculations in the vector space. It uses the term-document matrix (TDM) to reduce its dimension.

Methodology

The dataset of these processes is four translations of Surah Al-Fatihah as defined in Table 1, Table 2, Table 3 and Table 4. The proposed method in this paper presents three main steps, Preprocessing, TF- IDF, and Cosine similarity. Figure 1., illustrates the flow of the preprocessing step.

Table 1: Surah Al-Fatihah translated by Abdullah Yusuf Ali

Translated by Abdullah Yusuf Ali (T1)
In the name of Allah Most Gracious Most Merciful.
Praise be to Allah the Cherisher and Sustainer of the worlds.
Most Gracious Most Merciful.
Master of the Day of Judgment.
Thee do we worship and Thine aid we seek.
Show us the straight way.
The way of those on whom thou hast bestowed thy Grace.
Those whose (portion) is not wrath and who go not astray

Table 2: Surah Al-Fatihah translated by Ibn Kathir

Translated by Ibn Kathir (T2)
In the name God, the all-Merciful, the all-Compassionate.
All praise and gratitude (Whoever gives these to whomever for whatever reason and In whatever way from the first day of creation until eternity) are for God, the Lord of the worlds.
The All-Merciful, the All-Compassionate.
The Master of the Day of Judgment.
You alone do we worship, and from You alone do we seek help.
Guide us to the Straight Path.
The Path of those whom You have favored, not of those who have incurred (Your) wrath (punishment and condemnation), nor of those who are astray.

Table 3: Surah Al-Fatihah translated by Naasir Al-sa'di

Translated by Abd al-Rahman Ibn Naasir Al-Sa'di (T3)
In the name of Allah, the Entirely Merciful, the Especially Merciful.
[All] praise is [due] to Allah, Lord of the worlds –
The Entirely Merciful, the Especially Merciful,
Sovereign of the Day of Recompense.
It is You we worship and You we ask for help.
Guide us to the straight path –
The path of those upon whom You have bestowed favor, not of those who have evoked [Your] anger or of those who are astray.

Table 4: Surah Al-Fatihah translated by Sayyid Abul Ala Maududi

Translated by Sayyid Abul Ala Maududi (T4)
In the name of Allah, the Compassionate, the Merciful.
Praise is only for Allah, the Lord of the Universe,
the All-Compassionate, the All-Merciful,
the Master of the Day of Judgment
Thee alone we worship and to Thee alone we pray for help
Show us the straight way,
the way of those whom Thou hast blessed; who have not incurred Thy wrath, nor gone astray.

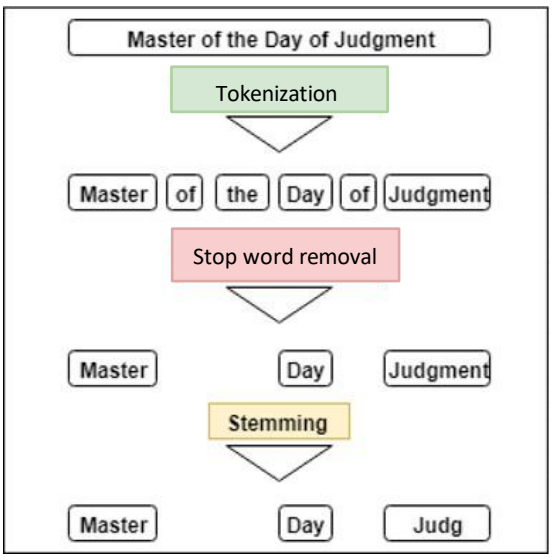

Figure 1: Preprocessing stages

Preprocessing

step 1: To splits documents into elements called tokens.
step 2: To removes all forms of punctuation and word that has no meaning or are not important. step 3: To remove the suffix at the end of a word to obtain the root word.

One chapter of Surah Al-Fatihah has seven verses for use as input it became documents. This preprocessing step required three stages. The first stage is Tokenization is the task of splitting the input text into very simple units, that called tokens, which generally correspond to words, numbers, and symbols, and are typically separated by white space in English (Diana et el., 2017). The next stage is stopword removal to remove all form of punctuation and word that has no meaning or are not important (Verma, 2014). Usually, stopword removes connecting words and prepositions. The last stage is stemming to remove the suffix at the end of a word to obtain the root word (Ferilli, 2014). In many documents, it can be found that several words are written in different forms, but it comes from the same root word. After completing preprocessing step next is to calculate TF-IDF weight.

In implementation, I use WEKA Workbench. WEKA is a data mining Workbench; it is a tool that I used for preprocessing the dataset using *the StringToWordVector* filter. The important setting parameters for this filter which is related to the preprocessing stages are
(1) *Stemmerweka.core.stemmers.LovinsStemmer* filter,
(2) Stopwords- handlerweka.core.stopwords.Rainbow and last filter
(3) Tokenizerweka.core.tokenizers.AlphabeticTokenizer filter.

TF-IDF

Term Frequency or TF is used to measure the frequency of the term in the document and Inverse Document Frequency or IDF to measure the importance of the term in all documents. With the combination of TF and IDF, we can calculate the TF-IDF form.

There are 5 steps of TF-IDF calculations as detailed below:
step 1 Calculate TF: find the number of times (frequency) the term appears in each document as given in Equation 1.

Equati$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$
step 2 *SUM of TF to find the number of times (frequency) the term appears in all documents.*
step 3 Calculate IDF: the value of DF divided by the total number of documents as defined in Equation 2.

Equati$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

step 4 Calculate Log(IDF): calculate the log base 10 of an IDF value.
*step 5 Calculate Weight of TF-IDF: TF-IDF score = TF * Log(IDF) as defined in Eqaution 3 and Equation 4.*

Equat$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

Equat$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Cosine Similarity

This study uses Cosine Similarity as a metric measuring and determines how similar the documents are regardless of their size as illustrated in Figure 2, in this case, is Quran dataset of four translations of Surah Al-Fatihah. Mathematically approach it measuring the cosine of the angle between two vectors projected in a multi-dimensional distance (Diana et al., 2017) as given in Equation 5.

tf-idf of document1 *(A)*	tf-idf of document2 *(B)*	A*B	A^2	B^2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
		SUM(A*B) *(C)*	SUM(A^2)	SUM(B^2)
			$\sqrt{SUM(A^2)}$ *(D)*	$\sqrt{SUM(B^2)}$ *(E)*
similarity = C / (D * E)				

Figure 2: Explained how to calculate the similarity

$$\text{cosine similarity} = S_C(A,B) := \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

There are 7 steps listed below for similarity calculation:
step 1: columns A and B are the tf-idf value of each document. step 2: multiply A by B.
step 3: calculate A power 2.
step 4: calculate B power 2.
*step 5: calculate the SUM of A*B, A^2, B^2. step 6: find the square root of A^2, B^2.*
*step 7: similarity = C / (D * E).*

Equation 5

Experimental Result

From the preprocessing method, the result is the dataset or terms which already been cleaned by tokenization, stopword removal, and stemming as presented in Table 5.

Table 5: The cleaned dataset.

Terms							
1	allah	14	show	27	exist	40	univers
2	grace	15	straight	28	lord	41	pra
3	merc	16	astra	29	owner	42	bles
4	nam	17	bestow	30	recompens	43	gon
5	cherishes	18	hast	31	alon	44	incur
6	prais	19	port	32	guid	45	entir
7	sustainer	20	thos	33	path	46	espec
8	world	21	thou	34	anger	47	du
9	judgm	22	thy	35	earn	48	soverei gn
10	master	23	wh	36	grant	49	ar
11	aid	24	whos	37	hav	50	evok
12	seek	25	wrath	38	compas	51	favor
13	worship	26	al	39	onl	52	day

Then, calculate the weight of TF-IDF as the normalized TF by counting the number of times (frequency) a word appears in a document, divided by the total number of words (frequency) in that document, and IDF is calculated the number of documents as logarithm in the corpus divided by the number of documents where the specific term appears (Diana et al., 2017). For example, the term *allah* appears 6 times. The total number of documents is 28. Hence the normalized term frequency is 6 / 28 = 4.66 andcalculate with log base 10 so the given below in Table 6 are the normalized term frequency for all the documents.

Table 6: TF-IDF calculation.

No	**Term**	**TF**			
		D7	**D14**	**D21**	**D28**
1	allah	0.00	0.00	0.00	0.00
2	merc	0.00	0.00	0.00	0.00
3	prais	0.00	0.00	0.00	0.00
4	world	0.00	0.00	0.00	0.00
5	judgm	0.00	0.00	0.00	0.00

6	master	0.00	0.00	0.00	0.00
7	seek	0.00	0.00	0.00	0.00
8	worship	0.00	0.00	0.00	0.00
9	straight	0.00	0.00	0.00	0.00
10	bestow	1.83	0.00	0.00	1.83
11	port	2.31	0.00	0.00	0.00
12	wrath	1.83	0.00	1.83	0.00
13	lord	0.00	0.00	0.00	0.00
14	owner	0.00	0.00	0.00	0.00
15	alon	0.00	0.00	0.00	0.00

Next, the result from similarity Table 7 which computes by using Cosine Similarity to match documents and calculated using the formula as in described in section 3.3, the result will show values of each document matching, the most similarity between document values is 1 and the most unmatched values are 0.

Table 7: Cosine similarity calculation.

Similarity						
T1 D1 **T2 D8**		**WD1**	**WD8**	**WD1*WD8**	**WD1^2**	**WD8^2**
1	allah	0.87	0.87	0.75	0.75	0.75
2	grac	1.07	1.07	1.14	1.14	1.14
3	merc	0.87	0.87	0.75	0.75	0.75
4	nam	1.35	1.35	1.82	1.82	1.82
5	cherishes	0.00	0.00	0.00	0.00	0.00
6	prais	0.00	0.00	0.00	0.00	0.00
7	sustainer	0.00	0.00	0.00	0.00	0.00
8	world	0.00	0.00	0.00	0.00	0.00
9	day	0.00	0.00	0.00	0.00	0.00
10	judgm	0.00	0.00	0.00	0.00	0.00
Sum				3.33	4047	6.67
SquertRoot					2.11	2.58
Similarity				0.61		

After analysing the similarity, the result show values of each document matching, for example, the similarity between Document 1 (D1) of Translation 1 (T1) and Document 10 (D10) of Translation 2 (T2) is the higher core 0.65 and the lower score is 0.09 between D2 and D9 as shown in Table 8. And the other similar result is shown in Table 9.

Table 8: Comparison between T1 and T2

	D8	**D9**	**D10**	**D11**	**D12**	**D13**	**D14**
D1	1	0.10	0.65				0.11
D2	0.09	0.18					
D3	0.65		1				0.17
D4				0.19			
D5					0.23		
D6						0.29	
D7	0.09		0.14				0.24

Table 9: Comparison between T2 and T3

	D15	**D16**	**D17**	**D18**	**D19**	**D20**	**D21**
D8	0.61	0.09	0.14				
D9	0.08	0.35	0.26				
D10	0.61		0.21				
D11				0.19			
D12					0.70		
D13						0.29	
D14							0.27

Finally, Table 10 shows summarize the similarity between translations, the difference ranges value showed that T2 and T4 is the most match with a 0.39 score and T3 with T4 is the lowest match with a 0.22 score, where T2 is Ibn Kathir translation and T4 is Sayyid Abul Ala Maududi translation.

Table 10: Summaries of similarity between four translations

Similarity	**T1**	**T2**	**T3**	**T4**
T1	1	0.34	0.38	0.24
T2	0.34	1	0.32	0.39
T3	0.38	0.32	1	0.22
T4	0.24	0.39	0.22	1

Conclusion

In this study, I introduced a method for context word documents using cosine similarity measurement to find the similarity value between the surah Al-Fatihah of 4 translators. The present work implements important pre-processing techniques namely, TF-IDF on surah Al-Fatihah of 4 translator datasets. And presents a method for analyzing Similar documents with the cosine technique similarity of the classification of documents to increase accessibility and able to suggest documents that are similar to each other, effectively comparing the similarities of surah Al- Fatihah to have a measure of the similarities of 4 translator bears many similarities in comparison. Cosine similarity is a very useful technique in comparing the meaning of surah Al- Fatihah. And comparison results were able to discover insights similarities link the meaning of the Qur'an surah al-Fatihah from 4 translators.

A comparison of the similarities is needed to continue studying with the next generation to learn about using cosine similarity or other techniques to measure the similarities of many surahs in the Qur'an to discover the percentage of similar meanings. So, with this technique, we get a lot of results for the study, even though they can't show the optimized results. But can eventually be brought back and able to present to people clearly.

References

Verma, T. (2014). Tokenization and Filtering Process in RapidMiner. *International Journal Application Information System – ISSN2249-0868 found. Computer Science FCS, New York, USA*, vol. 7, no. 2, pp. 16-18.

Ferilli, S., F. Esposito., & D. Grieco. (2014). Automatic learning of linguistic resources for stopword removal and stemming from text. *Procedia Computer Science*, vol. 38, no. C, pp. 116-123.

Matt, S. (2016) Introduction to the Quran. https://carm.org/quran-introduction. 24 April 2019.

Jannah, F. M., (2019) The Meaning of Surah 01 Al-Fatihah Opener. Audible Australia Pty Ltd. https://www.audible.com.au/pd/The-Meaning-of-Surah-01-Al-Fatihah-Opener-La-Apertura-From-Holy-Quran-Bilingual-Edition-English-Spanish-Audiobook/B07K1MSBHN.

Kidwai, R. (1987). Translating the untranslatable: a survey ofenglish translations of the quran. The Muslim World BookReview, vol. 7, no. 4, pp. 66–71.

Nassimi, D. M. (2008). A thematic comparative review of some en-glish translations of the qur'an. Ph.D. dissertation, Universityof Birmingham.

Mohammed, K. (2005). Assessing english translations of the qur'an. Middle East Quarterly.

Husain, B. (1931).The Holy Quran: A Translation with Com-mentary: According to Shia Traditions and Principles. Moayyedul-Uloom Association, Madrasatul Waizeen.

Robinson, N. "Sectarian and ideological bias in muslim trans-lations of the qur'an," Islam and Christian-Muslim Relations,vol. 8, no. 3, pp. 261–278, 1997.

Al-Azzawi, Q. O., & Alwan, Z. H. (2013). Translation assessmentof three english translations of words of patience in someselected verses. Journal of Advanced Social Research, vol. 3,no. 6.

Jurafsky D., and Martin J. H. (2008). Speech and Language Processing. Pearson Prentice Hall.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas G. W., & R. A. Harshman. (1990). Indexing by latent semantic analysi. Journal of the American Society for Information Science,vol. 41, no. 6, pp. 391–407.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman R. (1998). Using latent semantic analysis to improve access to textual information. In Proceedings of the SIGCHI *conference on Human factors in computing systems*. ACM. , pp. 281–285.

Diana, M., Kalina, B., & Isabelle, A. (2017). Natural Language Processing for the Semantic Web. *Synthesis Lectures On The Semantic Web:Theory And Technology – Issn 2160-4711* , pp. 12.