Analisis dan Implementasi *Similarity* dengan *Monolingual Alignment* pada Kisah Nabi Musa dalam Kitab Agama Islam (Al-Qur'an) dan Kitab Agama Kristen (AlKitab)

Analysis and Implementation Similarity using Monolingual Alignment for Prophet Musa in Holy Book of Islam (Al-Qur'an) and Holy Book of Cristian(Bible)

Wahyu Purbaningrum[1], DR.Moch. Arif Bijaksana,Ir.,M.Tech[2], Said Al Faraby,S.T.,M.Sc[3].

Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom
wpurbaningrum42@gmail.com, arifbijaksana@telkomuniversity.ac.id, said.al.faraby@gmail.com

Abstrak

Dalam kitab agama, tertulis beberapa kisah nabi. Kisah nabi yang tertulis di satu kitab agama tertentu, beberapa juga tertulis di kitab agama lain serta memiliki kesamaan dan kemiripan. Sebagai contoh adalah kisah Nabi Musa antara kitab agama Islam yaitu Al-Qur'an dan kitab agama Kristen yaitu Injil atau yang sekarang sering disebut dengan Alkitab. Adanya kemiripan dan kesamaan dari beberapa kisah nabi yang tertulis dalam kedua kitab tersebut telah banyak menjadi topik bahasan ditengah masyarakat. Oleh karena itu, diperlukan pengetahuan untuk menimbang dan melihat sisi kesamaan dan kemiripan kisah nabi yang termuat antara kedua kitab tersebut untuk mengetahui persamaan dan perbedaan kisah nabi yang tertulis didalamnya. Dengan menerapkan konsep kesamaan semantik antar teks, *monolingual alignment* dan *natural languange processing* akan dilakukan perbandingan kisah antara kedua kitab. Dalam penelitian ini, digunakan Kisah Nabi Musa sebagai dataset. Data masukan dalam tugas akhir ini adalah berupa pasangan ayat AlQur'an dan Alkitab mengenai kisah Nabi Musa dengan tingkat kemiripan berbeda-beda dan ada pula ayat yang tidak berpasangan dikarenakan tidak terdapat di dalam kitab yang berlawanan. Dataset didapatkan dari kisah Nabi Musa dari ayat-ayat AlQur'an dengan merujuk pada buku *Kisah Nabi Tafsir Ibnu Katsir* dan kisah Nabi Musa di Alkitab (buku Keluaran). Hasil dari penelitian adalah berupa nilai (angka) kemiripan dari penerapan meto*de monolingual alignment* dan kemudian dibandingkan dengan *gold standard* yang telah dibuat yaitu kemiripan pasangan ayat yang dibangun secara manual dengan intuisi manusia yaitu sebesar. Dari penelitian yang telah dilakukan diperoleh nilai korelasi sebesar 0.8164 untuk perhitungan dengan *single proportion* dan 0.8167 untuk perhitungan dengan *separate proportion*. Sehingga, dengan hasil penelitian yang didapatkan diketahui bahwa kisah Nabi Musa antara Al-Qur'an dan Alkitab teridentifikasi memiliki *similarity* atau kemiripan.

Kata kunci : kisah, Nabi Musa, kesamaan, kemiripan, Al-Qur'an, Alkitab, *monolingual alignment*, *gold standard*

Abstract

Known that each religion has a holy book. In the holy book was written the stories of prophets and messengers. Often found a story in one particular holy book was also written in another holy book from other religion and they both have similarity in certain aspects. As an example is a story of Moses, the prophet and messenger of Muslims which was written in The Quran, is also written in The Gospel, Christian's holy book, or known as The Bible. The similarities in the both holy books have become topics in the society. Therefore it takes knowledge to see and find out the similarities and the differences of the stories in both holy books. It could be implement by applying the concept of semantic textual similarity, monolingual alignment and natural languange processing, to compare stories from both holy books. In this research, Moses's stories is used as the dataset. The input data of this research is The Quran verses and The Bible verses about Moses stories. The dataset consist of unpaired verses and paired verses depend on the Dataset are obtained from the stories of Moses in The Quran verses, referring to the book Kisah Nabi Tafsir Ibnu Katsir, and the stories of Moses in The Bible verses, referring to The Exodus. The output of this research is the values (numbers) of similarity from the implementation of monolingual alignment method. The value then compared by the gold standard which is obtained manually by comparing the similarity rely on human intuition. In this research are obtained the correlation value of 0.8164 with single proportion and 0.8167 with separate proportion.

Keywords : *stories, Moses, similarity, monolingual alignment, gold standard*

1. Pendahuluan

Setiap agama memiliki kitab panduan masing-masing yang mengatur segala kehidupan penganutnya. Ajaran agama yang telah ada sekarang sebagian besar merupakan hasil perjalanan dan pengamalam para nabi terdahulu. Kisah nabi yang tertulis dalam kitab suatu agama, beberapa juga tertulis di kitab agama lain dengan penggambaran yang berbeda. Sebagai contoh kisah Nabi Musa dalam agama Islam yang tertulis di Al-Qur'an dan dalam agama Kristen yang tertulis di Alkitab yang memiliki beberapa kemiripan dan kesamaan. Kemiripan kisah nabi yang terkandung dalam Al-Qur'an dan Alkitab telah banyak jadi bahan perbincangan dan telah banyak sumber yang memaparkan hal tersebut. Untuk mengetahui kemiripan kisah nabi ini, harus membaca dan memahami ayat-ayat dalam Al-Qur'an dan Alkitab. Selain itu, perlu mengumpulkan ayatayat Al-Qur'an yang mengisahkan perjalanan nabi dikarenakan keberadaan ayatayat mengenai kisah nabi yang tidak terurut di dalam Al-Qur'an, berbeda dengan Alkitab yang susunan ayat mengenai kisah nabi telah terurut dan runtut berdasarkan urutan waktu. Permasalahan lain adanya penggalan kisah yang dijelaskan secara rinci didalam Al-Qur'an tetapi di Alkitab hanya dijelaskan secara singkat, dan sebaliknya. Sebagai contoh kisah ketika Fir'aun ingin membunuh Nabi Musa setelah mengetahui Nabi Musa membunuh kaumnya hingga Nabi Musa melarikan diri ke Madyan, didalam Al-Qur'an dikisahkan dalam tiga ayat sedangkan dalam Alkitab hanya satu ayat dan dari segi cerita lebih rinci di Al-Qur'an dari pada Alkitab. Oleh karena itu, diperlukan pengetahuan untuk menimbang dan melihat sisi kesamaan dan kemiripan kisah nabi dalam kedua kitab untuk mengetahui perbedaan dan persamaan isi dari kisah nabi tersebut. Dengan menerapkan konsep dan teori penambangan teks, *natural language processing*, dan metode *monolingual alignment* akan dilakukan perbandingan kisah Nabi Musa dari kedua kitab yaitu AlQur'an dan Alkitab.

Monolingual alignment merupakan salah satu metode untuk mengukur kemiripan dari sepasang kalimat [1]. *Monolingual alignment* melakukan identifikasi pada kata-kata atau frasa dalam dua buah kalimat untuk mengetahui kesamaan dari segi stuktur kata dan makna kedua kalimat tersebut. Pada tugas akhir ini menerapkan metode *monolingual alignment* karena metode tersebut telah cukup banyak digunakan dalam penelitian untuk mencari kesamaan dan kemiripan dari sepasangan kalimat maupun teks. Metode ini telah cukup banyak digunakan dalam kompetisi bidang STS (*Semantics Textual Similarity*) yaitu kompetisi SemEval. STS (*Semantics Textual Similarity*) merupakan konsep yang dapat dilakukan untuk pengukuran kesamaan makna antar dua buah data teks [7]. Dari 1 penelitian SemEval, metode ini merupakan metode yang sederhana dengan performansi yang paling baik dalam salah satu running di task STS pada SemEval 2014 dan SemEval 2015. Dataset yang digunakan dalam penelitian ini adalah berupa pasangan ayat Al-Qur'an dan Alkitab mengenai kisah Nabi Musa. Sistem yang dibangun dalam penelitian ini akan menghitung nilai kemiripan pasangan ayat. Untuk menghitung akurasi dalam penelitian ini, nilai kemiripan antar pasangan ayat dari sistem akan dibandingkan dengan nilai gold standard yang merupan nilai kemiripan secara manual dengan intuisi manusia.

2. Kajian Teori

2.1. Al-Qur'an

Al-Qur'an menurut bahasa memiliki arti bacaan atau sesuatu yang harus dibaca , djipelajari, sedangkan menurut istilah, Al-Qur'an adalah firman Allah yang diturunkan kepada Nabi Muhammad melalui perantara malaikat Jibril yang merupakan mukjizat dan menggunakan bahasa Arab [2]. Al-Qur'an terdiri dari 30 juz, 114 surat dan 6236 ayat yang berisikan tentang petunjuk dan pedoman bagi kehidupan umas Islam [2]. Dalam Al-Qur'an terdapat kisah-kisah Nabi dan Rasul. Sebagai contoh kisah Nabi Msa yang kisahnya paling banyak tertulisa dalam Al-Qur'an dan tertuang dalam 28 surat akan tetapi penggalan yang berkaitan dengan sejarah dan semata-mata kisah, seperti catatan tanggal lahir dan wafatnya tidak disebutkan dalam AL-Qur'an [3].

2.2. Alkitab

Alkitab adalah kitab suci bagi umat Kristen. Alkitab merupakan pedoman bagi umat Kristen dalam menjalani kehidupannya. Alkitab terdiri dari 66 kitab yang merupakan 39 kitab dari perjanjian lama dan 27 kitab dari perjanjian baru [4].

2.3. Korpus

Korpus (corpus) merupakan kumpulan dari beberapa teks sebagai sumber penelitian bahasa dan sastra syarat kumpulan teks tersebut digunakan sebagai objek dari penelitian bahasa dan sastra. Korpus terdiri dari dokumen teks dalam jumlah yang cukup banyak, bisa sampai jutaan dokumen.Teks-teks dalam korpus disusun dengan sistematis (Nesselhauf, 2005) untuk memudahkan pengelolaan.

2.4. *Text Mining*

Text Mining adalah konsep dan teknik untuk mencari pola dari sumber data teks, seperti dokumen Word, PDF, Excel, XML, kutipan teks, dll. Text mining merupakan proses penggalian informasi dari data dalam jumlah besar yang bermanfaat untuk tujuan tertentu [5]. Untuk mendapatkan informasi yang diinginkan, terdapat beberapa tahapan dalam penambangan teks untuk mempersiapkan teks agar lebih mudah diproses oleh sistem. Tahapan ini disebut *text prepocessing* yang terdiri dari *case folding* (proses penyamaan *case* dalam dokumen yang mengubah

huruf kapital menjadi hurup kecil), *tokenizing* (pemotongan teks menjadi bagian-bagian tertentu), *stopword removal* (penghilangan kata-kata yang tidak berkontribusi banyak pada isi dokumen) dan *lemmatization* (pengembalian kata-kata ke bentuk kata dasar).

2.5. *Semantic Textual Similarity* (STS)

Semantic Textual Similarity adalah konsep untuk mengukur kesetaraan semantic dari beberapa teks yang mirip. STS menilai seberapa mirip dua segmen teks [6].

2.6. Monolingual Alignment

Monolingual alignment merupakan metode *alignment* yang diimplementasikan untuk mencari kesamaan dan kemiripan semantik antar teks dalam satu bahasa atau dalam bahasa yang sama.

2.7. *Feature Extraction*

2.7.1. Word Similarity

Word Similarity mengidentifikasi kata-kata yang identik dan memiliki makna yang serupa. *Word Similarity* terdiri dari dua bagian yaitu *identical words* dan PPDB. Pada *word similarity* dilakukan tiga tahapan pengecekan kesamaan dan kemiripan kata, yang diantaranya adalah :

1. Jika saat *alignment* terdapat dua kata atau frasa yang identik pada kedua kalimat input, maka akan diberikan skor 1.
2. Jika saat alignment terdapat dua kata atau frasa yang memiliki kesamaan makna berdasarkan informasi dari korpus PPDB 1.0 XXXL maka akan diberikan skor 0.9.
3. Jika saat alignment, kata atau frasa tidak sama maka akan diberikan nilai 0. [7]

2.7.1.1. Align Identical Words

Fitur *identical words* pada *alignment* adala fitur yang mendeteksi kesamaan identik suatu kata. Kata dikatakan identik dengan kata lain apabila susunan huruf pada kata tersebut sama dan tidak dipengaruhi dengan ada atau tidaknya huruf kapital.

2.7.1.2. Align PPDB

Align PPDB mengidentifikasi kata-kata yang memiliki penulisan berbeda dengan makna yang sama. Berbeda dengan sinonim, jika sinonim hanya memeriksa kesamaan makna dari sepasang kata tunggal, sedangkan untuk PPDB ruang lingkupnya lebih luas, bisa berupa akronim, sinonim kata, frasa kata.

2.7.2. *Contextual Similarity*

Contextual similarity mengidentifikasi kemiripan dan kesamaan antara dua kalimat berdasarkan konteks kata-kata yang menyusun kalimat tersebut [6]. Adapun fitur-fitur *alignment* yang digunakan untuk *contextual similarity* adalah *synonym*, *named entities* dan *word sequences*.

2.7.2.1. Align Synonym

Align synonym adalah fitur alignment yang mengidentifikasi sinonim atau persamaan kata dari segi makna. Align *synonym* ini mengacu pada kamus kata tertentu sebagai contoh adalah *wordNet*.

2.7.2.2. Align Named Entities

Named Entities adalah kata atau frase yang menjelaskan mengenai nama orang, organisasi maupun tempat [12]. Fitur *align named entities* adalah fitur yang mengidentifikasi kata yang menunjukkan nama orang, organisasi dan tempat terhadap akronim atau sebutan lain bagi kata tersebut.

2.7.2.3. Align Word Sequences

Fitur *align word sequences* adalah fitur yang mengidentifikasi pasangan kata yang memiliki minimal dua urutan kata yang sama.

2.8. Perhitungan *Semantic Similarity*

Terdapat dua cara untuk melakukan perhitungan *semantic similarity*, yaitu *single proportion* atau *proporsi* tunggal dan *separate proportion* atau proporsi terpisah. *Single proportion* merupakan perhitungan yang tidak memperhatikan proporsi kata yang align pada suatu kalimat terhadap jumlah konten kata pada kalimat *pasangannya*. Adapun rumus dari *single proportion* adalah seperti pada Persamaan 2.1 berikut :

$$sts\left(S^{(1)}, S^{(2)}\right) = \frac{n_c^a\left(S^{(1)}\right) + \left(n_c^a S^{(2)}\right)}{n_c(S^{(1)}) + n_c(S^{(2)})} \quad (2.1)$$

dimana $n_c\left(S^{(i)}\right)$ dan $n_c^a\left(S^{(i)}\right)$ adalah jumlah *content words* dan jumlah *align content word* dalam $S^{(i)}$ [8].

Separate proportion merupakan perhitungan dengan memperhatikan proporsi kata yang align pada suatu kalimat terhadap jumlah konten kata pada kalimat pasangannya. Adapun rumus perhitungan untuk *separate proportion* adalah seperti pada Persamaan 2.2 berikut :

$$prop_{Al}^{(1)} = \frac{\left|\{i : [\exists j : (i,j) \in Al] \text{and } w_i^{(1)} \in C\}\right|}{\left|\{i : \ w_i^{(1)} \in C\}\right|} \quad (2.2)$$

di mana C adalah himpunan semua kata-kata konten dan Al adalah kata yang teridentifikasi align pada ayat kedua terhadap ayat pertama [11]. Sedangkan proporsi pada ayat kedua juga dapat dihitung dengan cara yang sama. Nilai *separate proportion* dari kedua ayat kemudian digabungkan dengan menggunakan *harmonic mean* [11]:

$$sim\ (S^{(1)}, S^{(2)}) = \frac{2 \times prop_{Al}^{(1)} \times prop_{Al}^{(2)}}{prop_{Al}^{(1)} + prop_{Al}^{(2)}} \quad (2.3)$$

Setelah diperoleh nilai *word silimilarity* dan *contextual similarity*, kemudian dilakukan perhitungan nilai kesamaan semantik dengan menggunakan perhitungan (2.4)

$$f(simW, simC) = 0.9 \times simW + 0.1 \times simC \quad (2.4)$$

dimana *simW* adalah hasil perhitungan *word similarity* dan *simC* merupakan hasil perhitungan *contextual similarity*. Bobot 0.9 dan 0.1 diturunkan secara empiris melalui pencarian grid dalam rentang nilai 0 hingga 1 untuk memaksimalkan kinerja *alignment* [7].

2.9. *WordNet*

WordNet adalah sebuah sistem leksikal yang dibangun secara manual dalam bahasa Inggris oleh George Miller dan koleganya di Cognitive Science Laboratory di Universitas Princeton (Fellbaum, 1998) dengan bagian dasar *synset* yang merupakan set sinonim dari sebuah konsep yang sama dipasangkan dengan penjelasannya seperti *gloseri* dari *synset*.

2.10. *Paraphrase Database* (PPDB)

Paraphrase database (PPDB) merupakan suatu basis data yang berisi kumpulan data parafrase. Parafrase adalah suatu kata yang memiliki makna yang sama, tetapi dalam bahasa yang berbeda atau tertulis dalam teks yang berbeda.

2.11. Evaluasi

2.11.1. Korelasi Pearson

Korelasi Pearson merupakan metode statistika yang digunakan untuk menghitung korelasi data atau menentukan apakah terdapat keterhubungan antar suatu set data. Rentang nilai korelasi pearson adalah antara -1 hingga 1. Perhitungan korelasi pearson adalah sebagai berikut :

$$Pearson = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \ \sum(Y - \bar{Y})^2}} \quad (2.5)$$

dimana X dan Y merupakan varibel nilai kesamaan semantik sistem dan *gold standard*. Dan $\bar{X}$ dan $\bar{Y}$ adalah rata-rata dari masing-masing variabel tersebut.

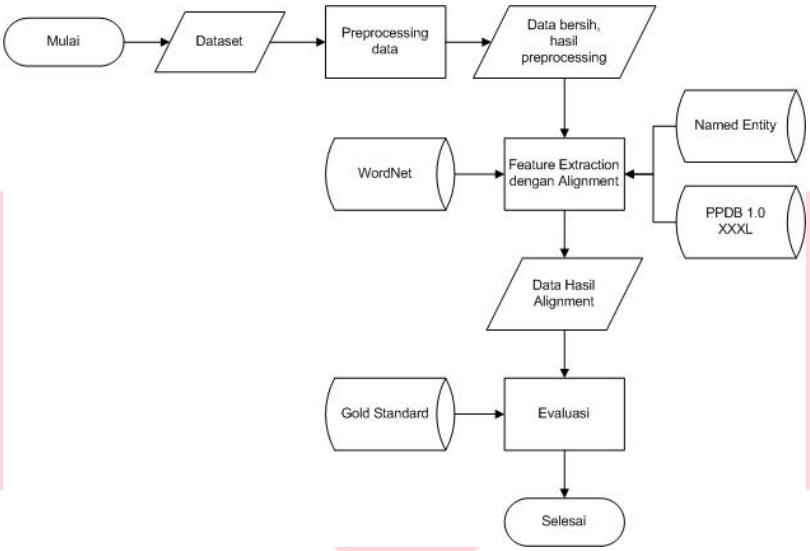
2.12. Gold Standard

Gold standard adalah penilaian dataset berdasarkan intuisi atau persepsi manusia. Semakin tinggi nilai yang didapatkan, maka semakin mirip isi dari pasangan ayat tersebut. Nilai ini akan digunakan sebagai pembanding hasil dari sistem.

3. Perancancangan Sistem

3.1. Gambaran Umum Sistem

Penelitian tugas akhir ini menggunakan metode monolingual alignment. Inputan yang digunakan adalah pasangan ayat Al-Qur'an dan Alkitab dalam bahasa Inggris. Tujuan dari sistem yang dibangun adalah untuk mendapatkan nilai kemiripan kisah Nabi Musa dalam Al-Qur'an dan Alkitab serta mengukur ketepatan metode yang digunakan dalam mengukur kesamaan dan kemiripan suatu teks atau kalimat yang sesuai dengan penilaian manusia. Gambaran umum dari sistem ini akan diilustrasikan pada Gambar 1.



Gambar 1: Gambaran Umum Sistem

Berikut penjelasan dari gambaran umum sistem pada Gambar 1 :

1. Sistem membaca dataset yang berupa pasangan ayat Al-Qur'an (terjemahan) dan Alkitab dalam bahasa Inggris.
2. Sistem melakukan *prepocessing* terhadap data, yaitu melakukan tokenisasi, *case folding, stopword removal, lemmarization*, dan menghapus *blank space*.
3. Setelah *prepocessing* didapatkan data bersih yang kemudian akan dilakukan beberapa *feature extraction* dengan *alignment*, yang diantaranya adalah *align identical word*, *align* PPDB, *align named entities*, *align synonym* dan *align word sequences.* Hasil dari *feature extraction* ini adalah nilai kesamaan atau nilai kemiripan dari masing-masing pasangan ayat dalam dataset. Nilai hasil dari masing-masing *feature extraction* ditampung atau disimpan dalam file dengan format .txt.
4. Sistem membaca nilai hasil *feature extraction* dan melakukan perhitungan nilai kesamaan atau kemiripan antara pasangan ayat dalam dataset.
5. Sistem melakukan evaluasi dengan cara membaca nilai *similarity* dan membaca nilai *gold standard* yang tersimpan dalam file dengan format .txt. Sistem melakukan evaluasi dengan menggunakan korelasi Pearson dengan rumus (2.3).

3.2. Pengumpulan Data

Data yang digunakan adalah ayat-ayat dari Al-Qur'an dan Alkitab mengenai kisah Nabi Musa. Ayat-ayat mengenai kisah Nabi Musa di Al-Qur'an belum terurut atau tersebar dalam surat-surat yang ada dalam Al-Qur'an sehingga dalam pengumpulan ayat-ayat Al-Qur'an mengacu pada buku Kisah Nabi Tafsir Ibnu Katsir. Buku ini merupakan buku kisah Nabi yang mengambil dari ayat-ayat Al-qur'an. Sedangkan untuk ayat-ayat Alkitab, didapatkan dari Kitab Keluaran yang merupakan kitab yang semua ayatnya membahas perjalanan Nabi Musa. Dataset terdiri dari 514 ayat Al-Qur'an dan 1191 ayat Alkitab (Kitab Keluaran). Dari seluruh ayat yang di dapatkan kemudian dipasangkan berdasarkan kemiripannya, sehingga didapatkan 95 pasang ayat dengan nilai kemiripan sangat tidak mirip hingga sangat mirip dan 1412 ayat tidak berpasangan. Pasangan ayat ini tidak selalu 1 ayat di AlQur'an berpasangan dengan 1 ayat di Alkitab, namun bisa dalam kondisi sejumlah ayat di Al-Qur'an berpasangan dengan sejumlah ayat di Alkitab. Sebagai contoh, 2 ayat di Al-Qur'an berpasangan dengan 1 ayat atau lebih di Alkitab dan sebaliknya. Dataset yang telah didapakan disimpan dalam bentuk file dengan format .txt, dengan tab sebagai pemisah antara pasangan ayat dan enter antar pasang ayat.

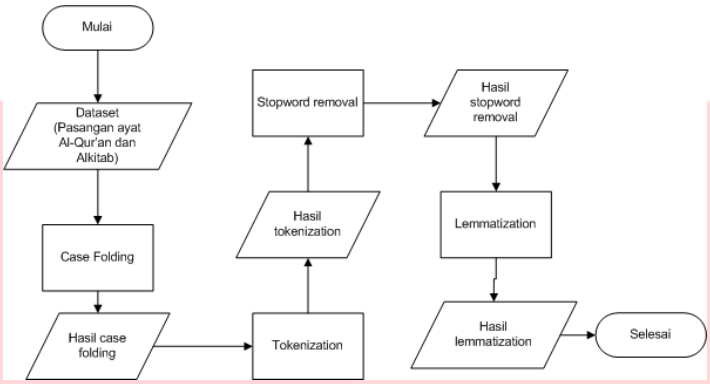
3.3. Pembuatan *Gold Standard*

Gold standard dibuat dengan cara menyebarkan kuisioner kepada 20 responden. Kuisioner berisikan seluruh pasangan ayat yang ada dalam dataset. Responden memberikan penilaian terhadap masing-masing pasangan ayat dengan range nilai antara 0 hingga 5. Dimana 0 adalah untuk pasangan ayat yang dianggap paling tidak mirip dan 5 adalah pasangan ayat yang dianggap paling mirip. Hasil penilaian 16 dari 20 responden ini kemudian dicari nilai rata-rata. Nilai rata-rata dari masingmasing ayat inilah yang akan dijadikan sebagai nilai *gold standard*.

3.4. Fungsionalitas Sistem

3.4.1. Preprocessing

Tahap awal yang dilakukan dalam sistem ini adalah *prepocessing*. Pada tahap ini, data akan dibersihkan dengan cara *case folding*, *tokenisasi*, *stopword removal*, *dan lemmatization*. Adapun tahapan *prepocessing* ini akan diilustrasikan pada Gambar 2.



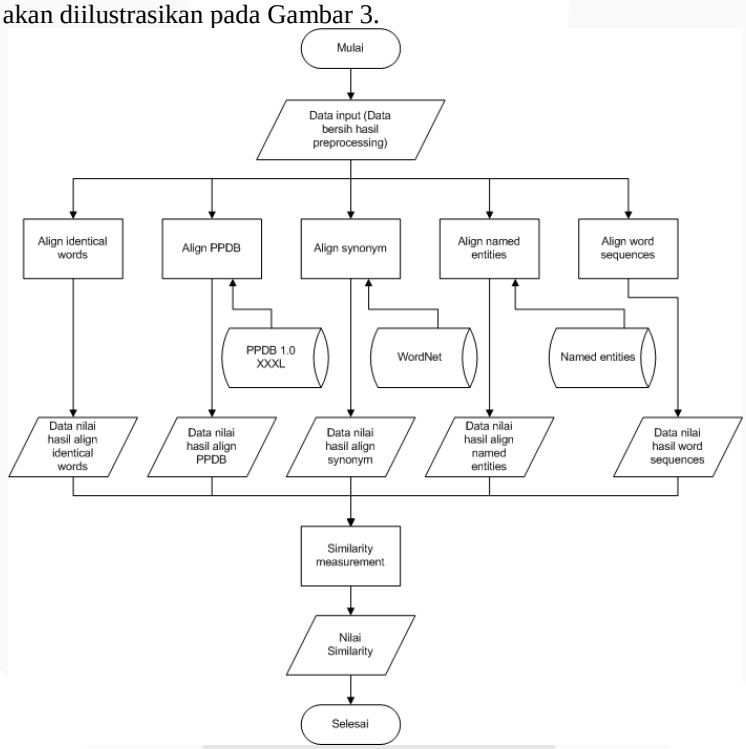
Gambar 2 : Tahap *Preprocessing*

Adapun penjelasan dari masing-masing tahapan *preprocessing* pada Gambar 2 adalah sebagai berikut :

a. *Case Folding* Dalam tahap ini, semua huruf kapital atau huruf besar diubah menjadi huruf kecil.
b. *Tokenization* Dalam tahap ini, teks yang berupa kalimat dan paragraf dipotong menjadi susunan per kata.
c. *Stopword Removal* Dalam tahap *stopword removal*, kata-kata yang dianggap tidak digunakan dan tidak berpengaruh terhadap isi dokumen akan dihapus atau dihilangkan.
d. *Lemmatization* Kata yang telah melewati proses *case folding*, *tokenization*, dan *stopword removal* akan dikembalikan kebentuk dasarnya.

3.4.2. *Feature Extraction* dengan *Alignment*

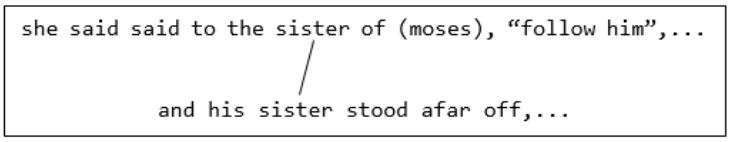
Sistem dibangun dengan menerapkan beberapa *feature extraction alignment* untuk mengetahui nilai kemiripan suatu teks atau suatu set data. Adapun *feature extraction alignment* yang diterapkan pada sistem yaitu *align identical words*, *align* PPDB, *align named entities*, *align synonym* dan *align word sequences*. Tahapan *feature extraction* ini akan diilustrasikan pada Gambar 3.



Gambar 3. *Feature Extraction* dengan *Alignment*

a. *Align Identical Words*

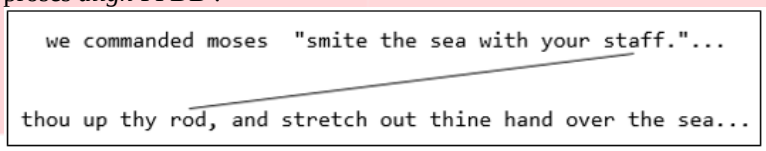
Setiap kata dalam kalimat hasil *prepocessing* dibandingkan dengan setiap kata dalam kalimat pasangannya. Sistem akan mengidentifikasi atau mencari kemiripan antar kata berdasarkan string atau hurufnya. Jika terdapat kata yang sama, maka akan dilakukan align dan diberi nilai 1 dan jika tidak sama maka akan diberi nilai 0. Berikut adalah contoh dari proses *identical words* :



Gambar 4 : *Align Identical Word*

b. *Align* PPDB

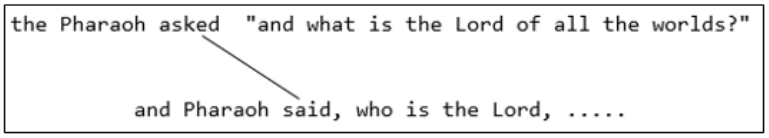
Sistem akan mengidentifikasi kata-kata atau frase yang memiliki makna yang sama. Pada tahap ini, sistem mengacu pada pharaphrase database yaitu PPDB vesi 1.0 XXXL. Setiap kata dalam inputan kalimat diperikasa ke dalam pharaphrase database, jika kata tersebut ada dalam database, maka akan dilakukan *align*. Berikut adalah contoh dari proses *align* PPDB :



Gambar 5 : *Align* PPDB

c. *Align Synonym*

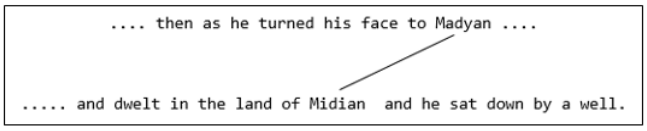
Sistem akan mengidentifikasi setiap kata pada data inputan pasangan ayat yang memiliki sinonim di kata dalam ayat pasangannya. Dalam tahap ini, sistem mengacu pada *corpus synset* yang terdapat pada wordNet. Dalam wordNet, sinonim antar kata memiliki nilai kemiripan antara 0 hingga 1. Dalam proses ini, sistem akan membaca token atau kata dari tiap-tiap pasangan ayat, kemudian sistem akan mengakses *corpus synset* wordNet untuk mengecek apakah antara token (kata) dari pasangan ayat memiliki sinonim. Tiap pasangan ayat bisa menghasilkan lebih dari satu nilai kemiripan, sehingga untuk mendapat nilai akhir *align synonym*, dilakukan perhitungan rata-rata nilai kemiripan kata dari tiap pasangan ayat. Berikut adalah contoh dari proses *align synonym* :



Gambar 6 : *Align Synonym*

d. *Align Named Entities*

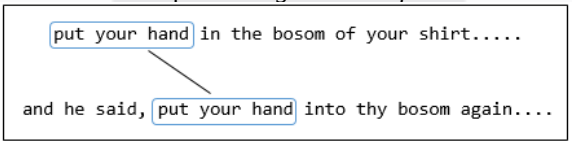
Sistem akan mengidentifikasi adanya kata yang memiliki makna entitas sama. Entitas yang dimaksud dapat berupa nama tempat, orang dan organisasi. Untuk identifikasi entitas, sistem mengacu pada korpus *named entities yang* telah dibangun oleh penulis dalam format .txt. Berikut adalah contoh dari proses align named entities :



Gambar 7 : *Align Named Entities*

e. *Align Word Sequences*

Dalam tahap ini, sistem akan mengidentifikasi dua atau lebih kata-kata identik yang ada pada masing-masing pasangan ayat. Berikut adalah contoh dari proses *align word sequences* :



Gambar 8 : *Align Word Sequences*

3.4.3. Perhitungan *Similarity*

Pada penelitian ini digunakan dua buah metode perhitungan nilai *similarity* yaitu perhitungan *single proportion* dan s*eparate proportion*. Sistem membaca hasil dari fitur alignment yang tersimpan kemudian melakukan perhitungan dengan menggunakan rumus (2.2), (2.3) dan (2.4)

3.4.4. Evaluasi

Setelah didapatkan semua nilai align, maka sistem akan melakukan evaluasi terhadap hasil perhitungan tersebut. Evaluasi dilakukan dengan membandingkan niali dari sistem dengan *gold strandard*. Evaluasi yang digunakan dalam sistem ini adalah dengan korelasi Pearson dengan melakukan perhitungan dengan rumus (2.5).

4. Pembahasan

4.1. Implementasi

Sesuai dengan rancangan sistem yang telah dijelaskan dalam bab tiga, sistem dibangun dengan bahasa pemrogramman Python 2.7. Dengan dataset pasangan ayat Al-Qur'an dan Alkitab mengenai kisah Nabi Musa serta data gold standard untuk menghasilkan nilai korelasi. Sistem dibangun dengan menerapkan kelima fitur *alignment* yang telah disebutkan dalam bab tiga, yaitu *align identical word, PPDB, synonym, named entities* dan *word sequences.*

4.2. Hasil Pengujian

Running program dilakukan dua kali yaitu dengan dataset keseluruhan ayat mengenai kisah Nabi Musa dan hanya ayat yang berpasangan. Adapun hasil dari pengujian yang dilakukan adalah :

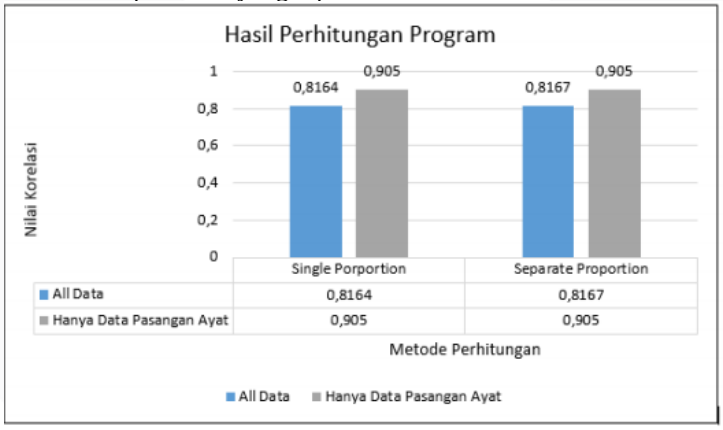
Dataset	Single Proportion	Separate Proportion
Seluruh data	0,8164	0,817
Hanya pasangan ayat	0,905	0,905

Selain itu, juga dilakukan running program hanya dengan melibatkan salah satu komponen penelitian yaitu *word similarity* yang teridiri *dari align identical words, align* PPDB dan *contextual similarity* yang terdiri dari *align synonym, align named entities, align word sequences*. Adapun hasil dari pengujian fitur adalah sebagai berikut :

Fitur Alignment	Single Proportion	Separate Proportion
Full Feature	0.8164	0.8167
Word Similarity	0.61104	0.61101
Contextual Similarity	0.8189	0.8204
(-) Align Identical Words	0.821	0.8218
(-) Align PPDB	0.8198	0.8204
(-) Align Synonym	0.6084	0.611
(-) Align Named Entitie	0.816	0.816
(-) Align Word Sequences	0.81	0.8104

4.3. Analisis Pengaruh Dataset

Berdasarkan hasil penelitian yang telah dilakukan dengan melakukan running dataset kisah Nabi Musa pada sistem tugas akhir maka didapatkan nilai korelasi yang menunjukan ketepatan metode yang digunakan dalam mendeteksi kemiripan dataset. Adapun hasil yang diperoleh akan diilustrasikan dalam Gambar 4.1 berikut :



Gambar 9 : Hasil penelitian

Berdasarakan Gambar 9, dapat diketahui bahwa hasil perhitungan dengan menggunakan *single proportion* adalah 0,8164 dan dengan menggunakan *separate proportion* sebesar 0,8167 untuk seluruh data. Sedangan untuk data hanya pasangan ayat, diperoleh hasil sebesar 0,905. Dengan demikian, dapat diketahui bahwa nilai korelasi dari data yang hanya pasangan ayat saja lebih tinggi dari pada hasil running dari seluruh data. Hal ini dapat terjadi karena jumlah data merupakan salah satu faktor yang dapat mempengaruhi nilai korelasi. Berdasarkan rumus Korelasi Pearson yang telah di jelaskan pada bab 2, diketahui bahwa faktor yang mempengaruhi nilai korelasi

adalah variabel X dan Y. Dimana X merupakan nilai *Gold Standard* dan Y merupakan nilai dari sistem. Dalam rumus Korelasi Pearson terdapat elemen yang melibatkan jumlah data, yaitu digunakan untuk menghitung nilai rata-rata dari masing-masing varibel. Sehingga apabila jumlah data lebih banyak, maka nilai korelasi yang dihasilkan dapat lebih sedikit, karena nilai pembaginya lebih besar.Selain itu, faktor banyaknya nilai *similarity* yang bernilai nol juga dapat memperngaruhi hasil korelasi menjadi lebih rendah.

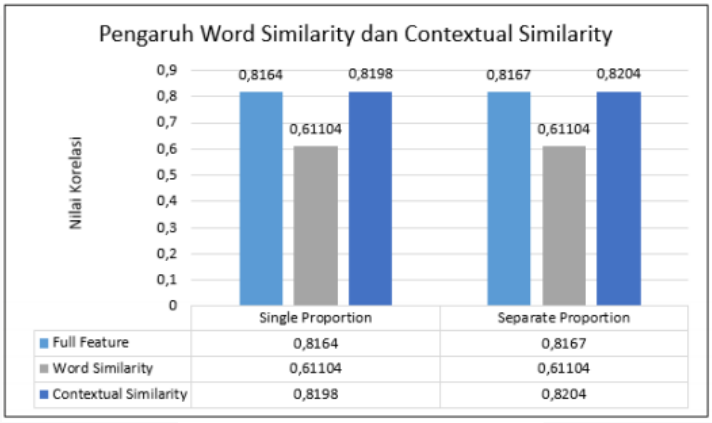
4.4. Analisis Pengaruh Metode Perhitungan *Similarity*

Berdasarakan dari hasil evalusi sistem yang diilustrasikan pada Gambar 9, diketahui bahwa hasil perhitungan dengan metode *separate proportion* lebih tinggi daripada perhitungan dengan menggunakan *single proportion*. Hal ini dikarenakan perhitungan *similarity* pasangan ayat dengan *separate proportion* menghasilkan nilai yang lebih mendekati *gold standard* daripada perhitungan *single proportion*. Kemiripan nilai yang dihasilkan dengan perhitungan *separate proportion* ini, dikarenakan perhitungan *separate proportion* memperhatikan proporsi kata yang teridentifikasi *align*. Berbeda dengan *single proportion* yang tidak mempertimbangkan proporsi kata yang teridentifikasi *align*, sehingga nilai kemiripan ayat yang dihasilkan bisa jauh dari *gold standard*. Sebagai contoh dari hasil fitur *synonym*, bisa terjadi banyak kata dalam bahasa inggris yang teridentifikasi sinonim, sehingga saat dilakukan perhitungan, jumlah kata yang teridentifikasi *align* dapat melebihi jumlah kata dalam konten (kata-kata dalam ayat). Dengan demikian, nilai kemiripan ayat yang dihasilkan bisa jauh dari *gold standard.*

4.5. Analsisis Pengaruh Fitur *Alignment*

4.5.1. Pengaruh *Word Similarity* dan *Contextual Similarity*

Pengujian dilakukan dengan menerapakan masing-masing bagian *alignment* yaitu *word similarity* yang terdiri dari dua buat fitur *alignment* yaitu *align identical words* dan *align* PPDB dan *contextual similarity* yang terdiri dari tiga buah fitur *alignment* yaitu *align synonym, align named entities* dan *align word sequnces*. Kedua bagian *alignment* ini diterapkan secara bergantian pada sistem untuk mengetahui pengaruh dari masing-masing bagian *alignment* terhadap nilai korelasi.

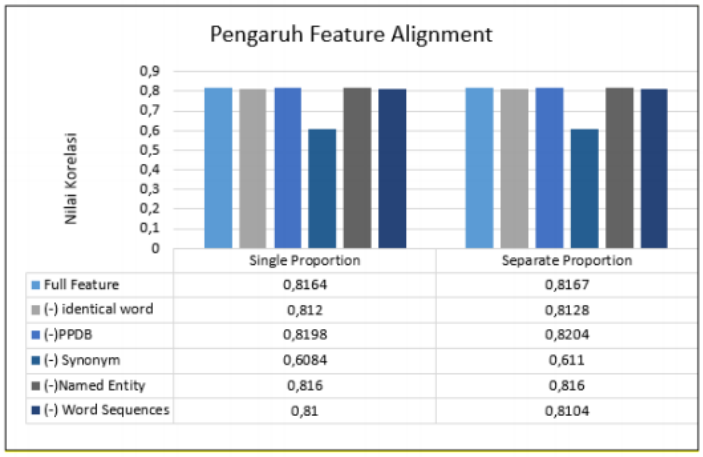


Gambar 10 : Pengaruh *Word Similarity* dan *Contextual Similarity*

Dari Gambar 9diketahui bahwa hasil perhitungan *similarity* dengan menerapkan komponen *alignment* dalam *contextual similarity* lebih tinggi daripada *word similarity*. Dengan demikian, *contextual similarity* lebih berpengaruh dan lebih efektif digunakan untuk pencarian *similarity text*. Hal ini dapat terjadi karena dalam *word similarity* hanya mengidentifikasi kata-kata dari segi susunan kata tanpa mempertimbangkan konteks kata dalam kalimat. Dapat dilihat dari fitur *alignment* yang menjadi komponen dari *word similarity*, yaitu *align identical words* dan *align* PPDB. Proses yang dilakukan oleh fitur *align identical words* dan *align* PPDB adalah mengidentifikasi kata-kata atau frasa yang identik. Walaupun *align* PPDB mempertimbangkan kesamaan makna dari frasa, tapi PPDB tidak melakukan pengecekan kesamaan makna dari tiap kata yang menyusun kalimat. *Align* PPDB hanya melakukan pengecekan terhadap frasa atau kata-kata tanpa mempertimbangankan konteks kata dalam kalimat. Selain itu hasil dari kedua fitur tersebut, jumlah kata yang teridentifikasi align dalam dataset sudah pasti dan tidak cukup banyak. Sebagai contoh kata *sister* pasti hanya akan teridentifikasi align jika terdapat kata sister di kalimat pasangannya. Berbeda dengan *contextual similarity* yang mengidetifikasi kata dengan mempertimbangkan konteks kalimat. Sebagai contoh *align synonym* yang melakukan pengecekan kesamaan makna dari setiap kata yang menyusun kalimat. Sehingga, memungkinkan adanya kata teridentifikasi memiliki kesamaan makna lebih dari satu kata lain yang menyusun kalimat. Dengan demikian, penilaian kesamaan dan kemiripan akan lebih rinci jika menerapkan *contextual similarity*.

4.5.2. Pengaruh Masing-masing Fitur Alignment

Pengujian dilakukan dengan cara menghilangkan atau tidak melibatkan satu persatu fitur alignment pada sistem secara bergantian. Hal ini bertujuan untuk mengetahui pengaruh dari masing-masing fitur terhadap nilai korelasi. Adapun hasil pengujian dari masing-masing fitur alignment ini akan diilustrasikan dalam Gambar 11.



Gambar 11. Pengaruh Masing-masing Fitur *Alignment*

Berdasarkan Gambar 9, dapat diketahui bahwa beberapa fitur *alignment* tidak terlalu berpengaruh terhadap nilai korelasi. *Align synonym* yang memiliki pengaruh paling besar terhadap nilai korelasi. Nilai korelasi yang dihasilkan tanpa melibatkan fitur *align synonym* mengalami penurunan paling banyak jika dibandingkan dengan fitur lain yaitu sebesar 0,208 untuk perhitungan dengan *single proportion* dan 0,2057 untuk perhitungan dengan *separate proportion*. Penurunan nilai korelasi yang cukup jauh ini terjadi karena, fitur *align synonym* melakukan pengecekan setiap kata dalam kalimat dengan mempertimbangkan konteks kalimat tersebut. Sehingga, kata-kata yang menurut penilaian intuisi manusia tidak mirip atau tidak sama, dalam sistem dapat teridentifikasi memiliki kesamaan makna berdasarkan kamus kata yang menjadi acuan sistem. Walaupun nilai kesamaan atau kemiripan yang tertera dalam kamus kata (korpus) tersebut sangatlah kecil, namun tetap memberikan pengaruh terhadap penilaian kemiripan pasangan ayat. Sehingga, apabila fitur *align synonym* tidak dilibatkan dalam sistem, maka nilai kemiripan kalimat yang didapatkan akan kecil karena hanya melibatkan fitur yang menilai dari segi identik kata, frasa, kesamaan entitas serta struktur kata yang terurut yang menghasilkan sedikit kata teridentifikasi align. Kata yang teridentifikasi *align* dengan fitur-fitur selain *align synonym* hanya sedikit, karena kata-kata dan frasa identik, entitas dan struktur kata yang terurut dalam dataset hanya sedikit. Sedangkan untuk fitur-fitur yang yang lain yaitu *align identical words*, *align* PPDB, *align named entities* dan *align word sequences*, tidak terlalu berpengaruh terhadap nilai korelasi. Hal ini dapat terjadi karena, jumlah kata *align* yang terdapat dalam dataset tidak cukup banyak. Sebagai contoh *align named entities*, pasangan ayat dataset yang mengandung kesamaan entitas hanya sedikit. Sehingga kata-kata yang teridentifikasi align juga hanya sedikit, yang mengakibatkan tidak berpengaruh besar terhadap nilai korelasi. Dalam dataset, hanya terdapat tiga pasang kata yang teridetifikasi *align*. Berikut kata-kata yang teridentifikasi align yang tedapat dalam dataset :

Kata1	Kata2
Madyan	Midian
God	Lord
Phrophet	Moses

Begitu pula dengan *align identical words, align* PPDB dan *align words sequences*. Jumlah kata yang teridentifikasi *align* untuk ketiga fitur *alignment* tersebut hanya sedikit. Selain itu juga dikarenakan, jumlah data yang memiliki pasangan ayat tidak cukup banyak yang mengakibatkan kata-kata yang teridentifikasi *align* untuk fitur *align identical words, align* PPDB, *align named entities* dan *align word sequences* juga tidaklah banyak.

5. Kesimpulan

5.1. Kesimpulan

Setelah melakukan evaluasi dari setiap hasil yang dihasilkan oleh penelitian ini, didapat beberapa kesimpulan yang diantaranya adalah :

1. Implementasi *monolingual alignment* untuk perhitungan similarity kisah Nabi Musa dengan dataset pasangan ayat Al-Qur'an dan Alkitab cukup tinggi yaitu dengan nilai korelasi 0,8164 untuk perhitungan *single proportion* dan 0,8167 untuk *separate proportion*
2. Perhitungan *similarity* dengan *separate proportion* lebih mendekati *gold standard.*

3. Fitur yang paling berpengaruh adalah *align synonym*, dengan penurunan nilai korelasi paling tinggi jika tanpa melibatkan fitur align synonym, yaitu sebesar 0.2

5.2. Saran

Untuk penelitian yang lebih baik dari sistem ini dalam mengidentifikasi kesamaan kalimat dengan menggunakan monolingual alignment maupun untuk pengembangan sistem, ada beberapa saran yang dapat dipertimbangkan untuk penelitian selanjutnya, yaitu :

1. Memperluas dataset yaitu mengenai kisah nabi yang lain
2. Memperbanyak isi *library Named Enitity*
3. Menambah fitur *alignment* yang dapat memperbesar nilai korelasi dan dapat mendeteksi kemiripan lebih spesifik
4. Mencoba menggunakan *library* lain atau kombinasi *library* untuk fitur *alignment* PPDB, *named entities*, *synonym* serta fitur lain yang memerlukan *library* untuk lebih meningkatkan performansi sistem dan agar dapat melengkapi *library* lain yang memungkinkan terdapat elemen yang belum tersimpan dalam *library* tersebut.

Daftar Pustaka

[1]. Hannah Bechara, Rohit Gupta, Liling Tan, Constantin Orasan, Ruslan Mitkov, and Josef van Genabith. Wolvesaar at semeval-2016 task 1: Replicating the success of monolingual word alignment and neural embeddings for semantic textual similarity. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 634–639, San Diego, California, June 2016. Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/S16-1096.

[2]. Ammi Nur Baits. Berapa jumlah ayat al-qur'an, 2013. URL: https:// konsultasisyariah.com/20876-berapa-jumlah-ayat-al-quran.html.

[3]. Anaksaleh.com. Cerita dalam al-qur'an, 2016. URL: http://www.anaksaleh. com/kisah-islami/cerita-al-quran.html.

[4].] Isa dan Islam. Alkitab-kitab suci agama kristen, 2015. URL: http://www. isadanislam.com/alkitab/alkitab-kitab-suci-umat-kristen.

[5]. Igg Adiwijaya. Text mining dan knowledge discovery. Komunitas Data mining Indonesia & Soft-omputing Indonesia, 2006.

[6].] Md Arafat Sultan, Steven Bethard, and Tamara Sumner. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. Transactions of the Association for Computational Linguistics, 2:219–230, 2014.

[7]. Md Arafat Sultan, Steven Bethard, and Tamara Sumner. Dls @ cu: Sentence similarity from word alignment. In SemEval@ COLING, pages 241– 246, 2014.

[8]. Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. A lightweight and high performance monolingual word aligner. In ACL (2), pages 702–707, 2013.

[9].] Hao Wu, Heyan Huang, and Wenpeng Lu. Bit at semeval-2016 task 1: Sentence similarity based on alignments and vector with the weight of information content. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 686–690, San Diego, California, June 2016. Association for Computational Linguistics. URL: http://www.aclweb. org/anthology/S16-1105.

[10]. Hans Lohninger. Pearson's correlation coefficient, 2016. URL: http://www. vias.org/tmdatanaleng/cc corr coeff.html.