Pembangunan Daftar Kata Terkait pada Kosa Kata Al-Qur'an Berdasarkan Kesamaan Distribusional

Fedy Fahron Guntara*[1], Moch. Arif Bijaksana[2], Arief Fatchul Huda[3]
[1,2]Universitas Telkom; Jl. Telekomunikasi no. 1, Terusan Buah Batu, Bandung, (022) 7565930
[3]UIN Sunan Gunung Djati; Jl. A. H. Nasution No. 105, Cipadung, Bandung, (022) 7800525
[1,2]Program Studi Informatika, Bandung
[3]Fakultas Sains dan Teknologi, Bandung
e-mail: *[1]fedyfahron@student.telkomuniversity.ac.id, [2]arifbijaksana@ telkomuniversity.ac.id, [3]afh@uinsgd.ac.id

Abstrak

Al-Qur'an adalah kitab suci umat Islam yang mengandung banyak kata di dalamnya. Hal tersebut membuat orang awam kesulitan untuk menemukan keterkaitan antar kata yang ada di dalam Al-Qur'an . Contohnya seperti kata مَّعْرُوف (perbuatan baik) yang memiliki keterkaitan dengan kata عَفَا (memaafkan) karena dalam Al-Qur'an kedua kata tersebut memiliki keterkaitan dalam makna yaitu memaafkan adalah salah satu perbuatan baik. Saat ini masih jarang ditemui kamus, ensiklopedia atau tesaurus kosa kata Al-Qur'an yang menjelaskan tentang keterkaitan antar kata dalam Al-Qur'an . Penelitian ini membahas tentang keterkaitan antar kata dalam Al-Qur'an dan kedepannya diharapkan dapat membantu dalam mencari keterkaitan antar ayat. Metode yang digunakan pada penelitian ini adalah metode dengan pendekatan kesamaan distribusional berbasis vektor Continuous Bag of Word (CBOW). Penggunaan metode CBOW menghasilkan nilai precision sebesar 98% berdasarkan dari hasil keluaran sistem dengan koreksi dari ahli bahasa.

Kata kunci— Al-Qur'an , *cosine-similarity*, *word2vec*, CBOW, *precision*

Abstract

The Quran is the Muslim holy book that contains many words in it. This makes it difficult for ordinary people to find connections between words in the Quran. For examples like the word مَّعْرُوف (good deeds) which have a connection with the word عَفَا (forgive) because in the Quran both words have a connection in the words of forgiveness is one of the good deeds. At present, there are still rarely found dictionaries, encyclopedias or thesaurus of the Quran vocabulary that explain the interrelationships of words in the Quran. This study discusses the interrelationship of words in the Quran and the future is expected to help in finding the interrelations between verses. The method used in this study is a method with a distribution-based equality approach based on Continuous Bag of Word (CBOW). The use of the CBOW method produces a precision value of 98% based on the results of the system output with the correction from linguists

Keywords— *Quran*, *cosine-similarity*, *word2vec*, CBOW, *precision*

1. PENDAHULUAN

Al-Qur'an adalah salah satu kitab suci atau buku umat Islam yang menjadi petunjuk setiap manusia. Sekitar 24% daripada populasi masyarakat dunia yang merupakan uamt Islam [1] membaca dan mempelajari tentang Al-Qur'an yang di dalamnya terdapat banyak pengetahuan atau informasi yang tersembunyi.

Al-Qur'an memiliki 30 juz, 114 surat dan lebih dari 6000 ayat yang terkandung di dalamnya [2]. Dari isi Al-Qur'an yang begitu banyak, membuat orang awam kesulitan untuk menemukan dan mempelajari tentang keterkaitan antar kata yang ada di dalam Al-Qur'an . Permasalahan tersebut dapat diatasi dengan adanya sebuah kamus, ensiklopedia atau tesaurus kosa kata Al-Qur'an dan sebagai salah satu kelengkapannya yaitu tiap entri kata memiliki keterkaitan dengan kata lain. Penelitian ini membahas tentang keterkaitan antar kata dalam Al-Qur'an dan untuk lebih lanjut diharapkan dapat membantu dalam mencari kemiripan antar ayat Al-Qur'an . Seperti contoh berikut ini: kata مَّعْرُوف (perbuatan baik) yang memiliki keterkaitan dengan kata عَفَا (memaafkan) karena dalam Al-Qur'an kedua kata memiliki keterkaitan yaitu salah satu perbuatan baik adalah memaafkan.

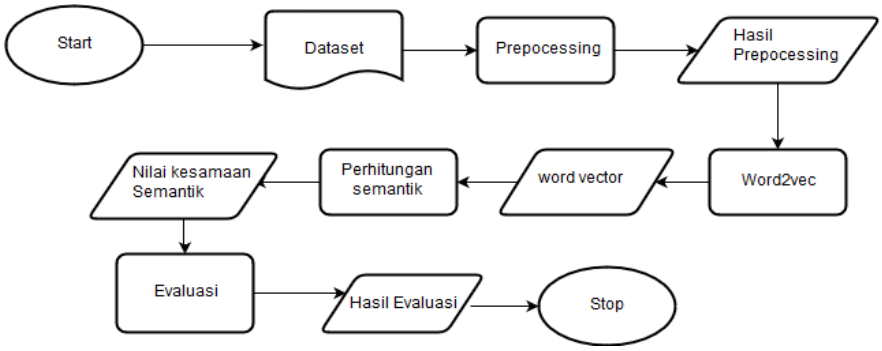
Keterkaitan dan kesamaan semantik berkaitan dengan bidang linguistik khususnya pada *Natural Language Processsing* (NLP) yang belakangan ini menjadi topik yang menarik dan banyak diteliti. Kesamaan semantik dan keterkaitan kata memiliki peran penting dalam beberapa *task* dari NLP dan beberapa bidang terkait seperti *text classification*, *document clustering*, *text summarization*, dan lain sebagainya [3]. Pada penelitian sebelumnya yang serupa hanya dilakukan analisis pada pencarian kata-kata yang mengandung atau menginterpretasikan manusia [4], kesamaan semantik antar teks terjemahan Inggrisnya [5]. Pada penelitian tersebut tidak terdapat penjelasan lebih mengenai keterkaitan antar kata dalam Al-Qur'an.

Melengkapi penelitian sebelumnya, maka penelitian ini menganalisis ke level keterkaitan antar kata dalam Al-Qur'an. Pembangunan sistem ini membutuhkan beberapa data seperti lema yang ada di dalam Al-Qur'an. Terdapat berbagai metode yang dapat digunakan untuk menganalisis keterkaitan antar kata. Salah satunya adalah menggunakan pendekatan kesamaan distribusional berbasis vektor, karena setiap kata memiliki ketergantungan terhadap jarak atau sudut yang dibangun antara vektor kata tersebut untuk dapat menghasilkan evaluasi kualitas representasi kata yang tinggi. Faktor tersebut yang menjadi alasan pengambilan basis vektor sebagai metode untuk mencari atau menghitung keterkaitan antar kata dalam Al-Qur'an. Keterkaitan antar kata tersebut dapat diketahui dari perhitungan nilai perbedaan sudut vektor kata-kata tersebut dengan menggunakan *cosine-similarity*. Sistem yang dibangun diharapkan dapat menghasilkan performansi yang baik berdasarkan nilai korelasi yang dihitung. Nilai korelasi yang dimaksud adalah nilai korelasi yang didapat dari perhitungan *precision* setelah dikoreksi oleh ahli bahasa.

2. METODE PENELITIAN

2.1 Gambaran Umum Sistem

Sistem yang dibangun bertujuan untuk menghasilkan himpunan kata yang memiliki keterkaitan dengan kata inputan melalui perhitungan *cosine-similarity* antara vektor kata yang didapatkan dari *word2vec*. Gambaran umum dari sistem yang akan dibangun dalam penelitian ini dapat dilihat pada Gambar 1. Pada Gambar 1 tersebut memvisualisasi tahapan-tahapan dalam penelitian ini mulai dari persiapan data, pembuatan model hingga evaluasi performansi.



Gambar 1. Gambaran Umum Sistem

2.2. Dataset

Dataset adalah sekumpulan data yang sudah diverifikasi kebenarannya dan dapat digunakan dalam penelitian sebagai sumber data yang valid. *Dataset* yang digunakan pada penelitian ini didapat dari situs *online* yaitu *corpus.quran.com* yang berisikan kumpulan lema yang terdapat pada Al-Qur'an . Kumpulan lema yang terdapat dalam *dataset* tersebut masih dalam bentuk kode *Buckwalter* sehingga perlu untuk diproses lebih lanjut.

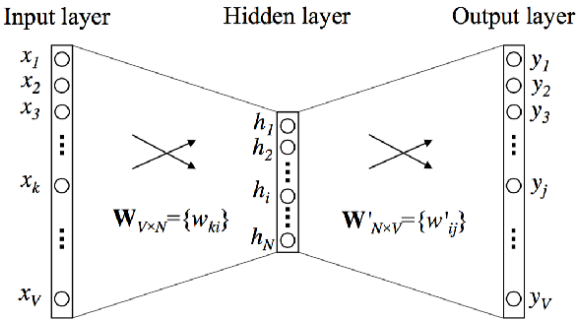
2.3 Preprocessing

Pada tahap *preprocessing* ini terdapat beberapa hal yang dilakukan diantaranya:

1. Pengambilan kata yang mempunyai lema
   Dataset yang digunakan masih terdapat beberapa kata yang tidak mempunyai lema sehingga perlu untuk dihilangkan.
2. Mengubah *Buckwalter* menjadi arab
   Dataset yang dipakai adalah *dataset* yang mengandung keseluruhan lema yang ada dalam Al-Qur'an dalam bentuk kode *Buckwalter* sehingga perlu untuk dilakukan *encode* dari *Buckwalter* ke arab.
3. Menyatukan lema berdasarkan ayatnya
   Proses ini melakukan penyatuan lema berdasarkan posisi ayatnya. Penyatuan lema ini digunakan pada saat proses ke dalam model *word2vec* sehingga dapat mengetahui setiap konteks dari lema tersebut dalam ayat.
4. Tokenisasi
   Ayat-ayat tersebut kemudian dilakukan tokenisasi sebelum diproses ke dalam model *word2vec* tokenisasi yang digunakan berdasarkan dari spasi.

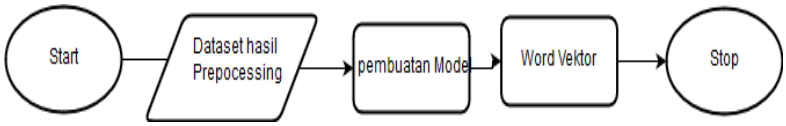
2.4 Word2vec

Word2vec adalah model yang digunakan untuk pemetaan kata menjadi vektor yang dibangun oleh Mikolov dari Google. Pada proses ini akan dilakukan perubahan kata-kata tersebut ke dalam vektor menggunakan metode *word2vec* tersebut, dengan tujuan untuk mencari kedekatan antar vektor yang satu dengan vektor yang lain untuk menemukan keterkaitan dan kedekatan dari kata tersebut. Metode yang digunakan adalah dengan memanfaatkan metode *Continuous Bag of Word* (CBOW). CBOW merupakan salah satu teknik dari *word2vec* yang menganalisa proyeksi vektor untuk memprediksi kata target berdasarkan dari konteksnya [6]. Untuk arsitektur dari metode CBOW sendiri dapat dilihat pada Gambar 2.



Gambar 2. Arsitektur CBOW [7]

Adapun urutan proses dari *word2vec* dapat dilihat pada Gambar 3.



Gambar 3. *Word2vec*

1. Data hasil dari proses *preprocessing* akan dignakan untuk proses pembentukan model dari *word2vec*.
2. Membangun konteks pasangan kata dari data korpus dengan berdasarkan jumlah *window size*, minimum frekuensi kata, *workers=8, alpha=0.22*. *Windows size* yang digunakan pada penelitian ini sebesar 5 dikarenakan pada penelitian sejenis dengan *windows size* 5 mengasilkan hasil yang optimal, minimum frekuensi kata yang digunakan sebesar 15 karena untuk mendapatkan lema-lema yang memiliki keterkaitan tinggi.
3. Setelah proses pembuatan model *word2vec* selesai, maka sistem menghasilkan vektor-vektor yang merepresentasikan lema dari *dataset*.

2.5 Perhitungan Keterkaitan dan Performansi

2.5.1 Perhitungan Keterkaitan Kata

Ketika kata direpresentasikan sebagai vektor istilah, kesamaan dua kata sesuai dengan korelasi antara vektor. Ini dikuantifikasi sebagai kosinus dari sudut antara vektor, yang disebut *cosine-similarity*. *Cosine-similarity* adalah salah satu perhitungan kedekatan vektor paling populer yang diterapkan pada dokumen teks, seperti dalam berbagai aplikasi pencarian informasi [8] dan pengelompokan juga [9].

Setelah melakukan perubahan kata-kata menjadi vektor dengan *word2vec* selanjutnya akan dihitung kedekatan antar kata-kata tersebut menggunakan rumus *cosine-similarity* [10] pada Persamaan 1.

$$\text{Sim}=\cos(\Theta) = \frac{\vec{A}\cdot\vec{B}}{|\vec{A}|\,x\,|\vec{B}|} \qquad (1)$$

2.5.2 Perhitungan Performansi

Setelah mendapatkan hasil kedekatan dari perhitungan *cosine-similarity* yang kemudian akan dilakukan perhitungan performansi dengan menghitung nilai *precision* hasil keluaran dari sistem dengan melakukan koreksi langsung ke ahli bahasa. *Precision* adalah rasio jumlah dokumen relevan yang ditemukan dengan total jumlah dokumen yang ditemukan oleh sistem [8]. Penggunaan *precision* pada penelitian ini untuk mengetahui tingkat ketepatan antara

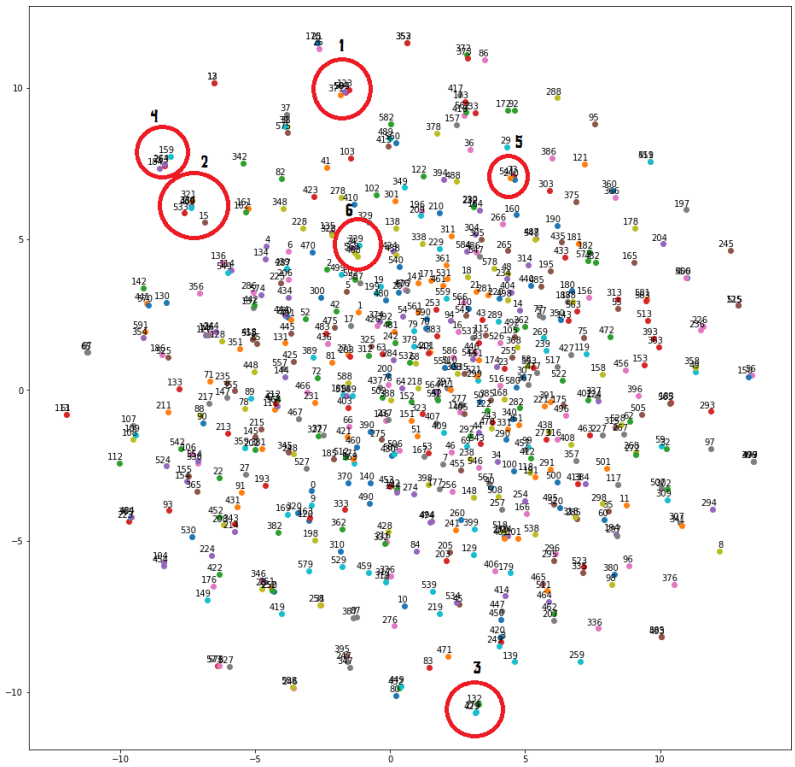
informasi yang sebenarnya dengan jawaban yang diberikan oleh sistem. Untuk menginterpretasikan angka *precision*, ditetapkan lima kategori atau kelas yaitu: presisi sangat tidak akurat, tidak akurat, cukup akurat, akurat dan sangat akurat. Sedangkan kriteria atau tolak ukur yang digunakan untuk menyatakan kategori tersebut ialah skala interval, dengan mencari selisih kemungkinan angka *precision* sangat akurat (1) dengan kemungkinan angka *precision* sangat tidak akurat (0) kemudian dibagi 5 sesuai kategori penilaian, riilnya adalah (1 - 0) : 5 = 0,20. Dengan demikian kelas interval dari kelima kategori interpretasi *precision* tersebut dapat dilihat pada Tabel 1.

Tabel 1. Kategori *Precision*

Persentase (%)	Keterangan
0-20	Sangat tidak akurat
21-40	Tidak akurat
41-60	Cukup akurat
61-80	Akurat
81-100	Sangat akurat

3. HASIL DAN PEMBAHASAN

Hasil dari penelitian ini adalah himpunan kata yang memiliki keterkaitan satu sama lain berdasarkan hasil perhitungan *cosine-similarity* antara vektor kata tersebut. Pada Gambar 4 terdapat visualisasi dari kata-kata yang terdapat didalam Al-Qur'an yang telah direpresentasikan menjadi 2 dimensi. Titik-titik yang terdapat pada Gambar 4 mewakili lema yang ada dalam Al-Qur'an berdasarkan dari nilai vektor yang dihasilkan dari proses *word2vec* dan untuk lingkaran yang terdapat pada Gambar 4 menggambarkan kumpulan kata yang mungkin memiliki keterkaitan.



Gambar 4. Visualisasi Vektor Kata

Contoh hasil dari keluaran sistem dengan masukan kata مَّعْرُوف dapat dilihat pada Gambar 5. Sistem akan mengeluarkan 10 kata yang memiliki nilai keterkaitan tinggi dengan kata masukan.

```
[('0.6529487371444702 ','مُنكَر'),
 ('0.5446025729179382 ','جُنَاح'),
 ('0.4986034631729126 ','نَهَىٰ'),
 ('0.4461979866027832 ','فَعَلَ'),
 ('0.444547176361084 ','بَعْض'),
 ('0.441963255405426 ','عَفَا'),
 ('0.43862223625183105 ','قِسْط'),
 ('0.42538732290267944 ','مُؤْمِنَٰت'),
 ('0.42476922273635864 ','مَتَٰع'),
 ('0.42437413334846497 ','سَهْر')]
```

Gambar 5. Contoh Keluaran Sistem

Visualisasi titik yang dilingkari pada Gambar 4 menjelaskan beberapa kata yang memiliki keterkaitan tinggi berdasarkan kedekatan antar vektor yang terbentuk. Kata-kata tersebut dapat dilihat di dalam Tabel 2.

Tabel 2. Perhitungan *Precision*

Nomor Titik	Kata	Keterangan	*True Positif*
1	مَّعْرُوف	Perbuatan baik	10
2	ذُرِّيَّة	Keturunan / anak	9
3	عَلِيم	Pengetahuan	10

4	مَرْيَم	Keluarga Maryam	10
5	أَذْنٌ	Izin Allah	10
6	عَلِم	Pengetahuan	10
Precision			59/60=0.98

Dari hasil pengujian dapat diketahui nilai *precision* untuk setiap kata uji. Perhitungan yang dihasilkan dari pengujian terdapat pada Tabel 2 yang menunjukkan angka *precision* sebesar 98%. Nilai *precision* yang didapatkan menunjukan tingkat ketepatan antara informasi yang sebenarnya dengan jawaban yang diberikan oleh sistem. Nilai *precision* pada hasil pengujian sebesar 98% termasuk dalam kategori sangat akurat.

4. KESIMPULAN

Dari hasil penelitian menghasilkan kesimpulan yaitu perhitungan *precision* menunjukan bahwa sistem yang dibangun memberikan tingkat ketepatan antara informasi yang sebenarnya dengan jawaban yang diberikan oleh sistem sebesar 98%. Pendekatan distribusional dan perhitungan *cosine-similarity* dapat menangani permasalahan keterkaitan antar kata dalam Al-Qur'an berdasarkan dari *precison* yang didapatkan.

5. SARAN

Berikut beberapa saran untuk penelitian tentang keterkaitan kosa kata Al-Qur'an:

1. Penelitian selanjutnya disarankan dapat mencoba menganalisis parameter yang digunakan pada model *word2vec*.
2. Dan tidak terlepas juga untuk menggunakan metode lain baik yang berbasis vektor atau metode yang tidak memanfaatkan vektor kata.
3. Hasil dari penelitian ini diharapkan dapat memberikan masukan dan menjadi acuan bagi peneliti selanjutnya untuk melakukan penelitian yang sama.

UCAPAN TERIMA KASIH

Dengan penuh rasa syukur kehadirat Allah SWT dan setelah itu tanpa menghilangkan rasa hormat yang mendalam, saya selaku penyusun dan penulis mengucapkan terima kasih yang sebesar-besarnya kepada Ustaz Nur Muttaqien selaku ahli bahasa Arab Al-Qur'an yang telah membantu penulis untuk menyelesaikan penelitian ini.

DAFTAR PUSTAKA

[1] A. H. Daud, Z. B. Othman, and N. A. Idris 2018, "Fahaman Syiah dan Keharmonian Agama Islam Di Malaysia: Perspektif Pendekatan Keselamatan," *J. Soc. Sci. Humanit.*, Vol. 13, No. 3, pp. 1–19.

[2] Sahabuddin, M. Q. Shihab, and Sahabuddin 2007, *Ensiklopedia Al-Qur'an: Kajian Kosakata*, Lentera Hati.

[3] R. Mihalcea, C. Corley, and C. Strapparava 2006, "Corpus-Based and knowledge-based measures of text semantic similarity," in *Proceedings of the National Conference on Artificial Intelligence*.

[4] A. W. Z. Nasution, M. A. Bijaksana, and S. Al Farab 2017, "Analisis dan Implementasi Perhitungan Semantics Similarity Pada Ayat Al-Quran Dengan Pendekatan Word Alignment Berdasarkan Support Vector Regression," *eProceedings Eng.*, Vol. 4, No. 2.

[5] M. M. Rani, M. A. Bijaksana, and S. Al Faraby 2017, "Analisis Dan Implementasi Kesamaan Semantik Antar Teks Menggunakan Pendekatan Alignment Dan Vektor Semantik Pada Terjemahan Alquran," *eProceedings Eng.*, Vol. 4, No. 2.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean 2013, "Efficient Estimation of Word Representations in Vector Space," *arXiv Prepr. arXiv1301.3781*.

[7] X. Rong 2014, "word2vec Parameter Learning Explained", *arXiv Prepr. arXiv1411.2738*.

[8] R. Baeza-Yates, B. Ribeiro-Neto, and Others 1999, *Modern Information Retrieval*, Vol. 463. ACM press New York.

[9] Aggarwal, Charu C., and ChengXiang Zhai, 2012, *eds. Mining Text Dat*, Springer Science & Business Media.

[10] A. Huang 2008, "Similarity Measures for Text Document Clustering," in *New Zealand Computer Science Research Student Conference, NZCSRSC 2008 - Proceedings*.