# Analisis dan Implementasi *Short Text Similarity* dengan Metode *Latent Semantic Analysis* Untuk Mengetahui Kesamaan Ayat al-Quran

## *Short Text Similarity Analysis and Implementation with Latent Semantic Analysis Method to Find al-Quran Verse Similarity*

**Mochamad Irfan Dary**[1]

[1] Prodi S1 Teknik Informatika, Fakultas Teknik, Universitas Telkom

**irfandary@gmail.com**[1]

**Abstrak**

**Salah satu cara untuk membantu memahami sebuah konsep adalah dengan mencari hubungan antara konsep tersebut dengan konsep yang lainnya. *Latent semantic analysis* (LSA) adalah salah satu penerapan dari hal tersebut. Dimana LSA dapat mencari nilai kesamaan antar konsep. LSA adalah suatu cara untuk mengekstraksi tulisan dan membuat representasi statistik dari tulisan tersebut dengan pendekatan dimensi rendah (*low rank approximation),* dimana digunakan dimensi yang kecil untuk mengetahui keseluruhan makna dokumen. Tetapi salah satu kekurangan LSA adalah tidak ada cara yang pasti untuk mengetahui berapa dimensi yang harus digunakan dan bagaimana signifikansi dari penggunaannya pada *short text* seperti ayat al-Quran. Dari hasil pengujian didapat bahwa dari berbagai dimensi yang digunakan, akurasi maksimum adalah 71% dan F-measure terbaik adalah 40%. Lebih baik dibandingkan *term document matrix* biasa (tanpa LSA) dengan akurasi 85% tetapi F-measure 0%. Temuan lainnya adalah jika dimensi yang dipilih terlalu rendah, maka nilainya semakin seragam dan bias sehingga nilainya kurang relevan. Selain itu ada dimensi keseimbangan, dimana dimulai dari dimensi tersebut, hasil *similarity* tidak akan terlalu jauh berbeda.**

**Kata Kunci: *latent semantic analysis*, *similarity*, al-Quran, *clustering*, matriks, *singular value decomposition, cosine similarity***

**Abstract**

**One way to help understand a concept is to find a relation between concept with other concepts. Latent semantic analysis (LSA) is an application of it. Where LSA can find similarity between concept. LSA is a method to extract text and create a statistical representation of the text with low rank approximation where low dimension are used to determine the overall meaning of the document. But one drawback of LSA is no sure way to know how many dimensions that should be used and how the significance of its use at short text like al-Quran. From the test results obtained from the various dimensions that are used, maximum accuracy is 71% and the best F-measure is 14%. Better than the usual term document matrix (without LSA) with an 85% of accuracy, but the 0% F-measure. Other findings is if the selected dimension is too low, then the value is more uniform and biased so that the value is less relevant. Additionally there is a dimension of balance, which starts from the dimensions, similarity results will not be too much different.**

**Keyword : Latent Semantic Analysis, Similarity, Al-Quran, Clustering, Matriks, Singular Value Decomposition, Cosine Similarity**

## 1. Pendahuluan

Al-Quran adalah salah satu kitab atau buku yang paling banyak dibaca dan dipelajari oleh manusia di seluruh dunia dan mempelajarinya adalah wajib bagi umat Islam. Al-Quran diklaim sebagai sumber data teks yang dipelajari oleh berbagai disiplin ilmu, karena di dalamnya terdapat banyak sekali cerita sejarah, metafora, dan deskripsi tersirat mengenai alam semesta dan segala isinya. Hal inilah menyebabkan penelitian mengenai al-Quran adalah penelitian yang bersifat kolaboratif dan mau tidak mau akan melibatkan banyak pihak dari disiplin ilmu yang beragam. Selain itu, setidaknya ada 1,5 milyar manusia yang diwajibkan untuk mempelajari al-Quran dan maknanya dan jumlah ini terus menerus bertambah, sementara untuk mempelajari makna al-Quran dan mengetahui keterkaitan antar ayat bukanlah sesuatu yang mudah dan cepat. Dimana semantik al-Quran cukup sulit untuk dimengerti dikarenakan maknanya yang kompleks. Untuk mengatasi masalah tersebut, dalam tugas akhir ini digunakan metode *latent semantic analysis* mencari nilai kesamaan pada ayat al-Quran, dalam kasus ini adalah surat Maryam. Metode ini menekankan pada pengamatan terhadap kata-kata kunci yang menjadi acuan penilaian tanpa memperhatikan karakter linguistiknya menggunakan proses-proses matematis yang menitikberatkan pada pengolahan matriks [1]. salah satu kekurangan LSA adalah tidak ada cara yang pasti untuk mengetahui berapa dimensi yang harus digunakan dan bagaimana signifikansi dari penggunaannya pada short text seperti ayat al-Quran.

## 2. Dasar Teori dan Perancangan

### 2.1. *Latent Semantic Analysis* (LSA)

*Latent semantic analysis* (LSA) adalah suatu cara untuk mengekstraksi tulisan dan membuat representasi statistik dari tulisan tersebut [2]. LSA adalah salah satu cara untuk mensimulasikan cara otak manusia menangkap representasi informasi dari sebuah tulisan. LSA menggunakan pendekatan matematis untuk mengetahui makna dari suatu teks dengan cara memanfaatkan fungsi statistik dari data yang berbasis korpus dengan pendekatan dimensi rendah (*low rank approximation)*. Konsep LSA ini akan mencari nilai kesamaan diantara 2 buah segmen teks tanpa memperhatikan susunan kata. LSA ini bisa menangkap asosiasi antara kata dengan artikel memanfaatkan pendekatan dekomposisi matriks yang didapat dari *term-matrix* dokumen [3]. Metode ini berpatokan pada prinsip dimana sebuah ide atau konsep yang ada dalam sebuah teks bisa diketahui cukup dengan kemunculan katanya saja, tidak perlu memperhatikan urutan kata-kata. Konteks kalimat bisa didapat dari

diksi yang digunakan, karena tiap kata kunci memiliki makna yang memiliki kaitan dengan dokumen dan dianggap sudah cukup merepresentasikan ide.

### *2.2. Singular Value Decomposition* (SVD)

SVD adalah cara dekomposisi matriks yang digunakan untuk mencari kesamaan antar segmen kata [2]**.** SVD adalah komponen pemrosesan yang mengkompresi informasi yang berkaitan dalam jumlah besar ke dalam ruang yang lebih kecil. Proses awal dari akan LSA merepresentasikan isi kata dalam matriks dua dimensi yang besar yang berisi *bag-of-words* dari surat Maryam dimana kolom merepresentasikan ayat, dan baris mewakili kata. Formula dari SVD adalah sebagai berikut :

$$A \approx \mathbf{USV^T} \qquad \mathbf{(1)}$$

Dimana :

A : matrix asal

U : orthonormal eigenvector dari $AA^T$

S : matriks diagonal

$V^T$ : transpose dari orthogonal matriks V

Dekomposisi ini memungkinkan dimensi matriks asal untuk dilakukan reduksi dimensi. Dengan proses reduksi dimensi terhadap perkalian matriks SVD, maka akan diperoleh penyederhanaan dan pembobotan dari matriks asal dengan mengambil sebagian besar dari struktur penting antara kata kunci dengan kalimatnya.

### *2.3. Cosine Similarity*

*Cosine similarity* adalah cara yang digunakan untuk membandingkan jarak dari 2 buah vektor. Dalam penelitian ini, nilai vektor didapat dari pemrosesan teks yang representasinya telah dirubah menjadi matriks. Dengan teks yang direpresentasikan ke dalam matriks dan diubah ke dalam matriks yang merupakan bentuk vektor, kita bisa mengetahui perbedaan sudut kosinus dari 2 buah vektor. Besar sudut inilah yang mengindikasikan besarnya perbedaan makna dalam teks.

Sementara formula untuk mengetahui nilai kosinus adalah sebagai berikut,

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (2)$$

Dimana : A: nilai vektor dari dokumen A

B : nilai vektor dari dokumen B

## 2.3. *F-measure Evaluation*

*F-measure* adalah pengukuran kualitas dari akurasi sebuah klasifikasi biner. Pengukuran ini menguji apakah akurasi yang dihasilkan dari sistem sesuai dengan klasifikasi natural manusia. Akurasi tidak selamanya mencerminkan kualitas dari sistem, karena sistem klasifikasi benar yang baik seharusnya dapat menangkap data yang minoritas. Pada kenyataannya hasil minoritas inilah yang dicari. Misalnya pada mesin pencari, data yang tidak berkaitan dengan *query* pencarian akan selalu lebih banyak dari pada yang berkaitan.pada *f-measure*, hasil evaluasi tidak hanya benar atau salah, melainkan diperluas menjadi berapa data benar yang diklasifikasi sebagai benar (*true positive*), berapa data salah yang dilabeli salah (*true negative*), berapa data benar yang dilabeli salah (*false negative*), dan berapa data yang salah dilabeli benar (*false negative)*. *F-measure* mempertimbangkan 2 hal, yaitu *recall* dan *precision.*

### 2.3.1. *Recall*

*Recall* mengindikasikan kuantitas dari sistem. *Recall* adalah perbandingan antara jumlah data yang relevan yang berhasil teridentifikasi oleh sistem dengan jumlah data relevan yang seharusnya teridentifikasi

$$\text{Recall} = \frac{tp}{tp + fn} \quad (3)$$

Dimana :

tp : *True Positive*

tn : *True Negative*

### 2.3.2. *Precision*

*Precision* menunjukkan kualitas dari klasifikasi sistem. Precision adalah persentase dari jumlah data yang relevan yang berhasil teridentifikasi oleh sistem terhadap jumlah data keseluruhan yang teridentifikasi

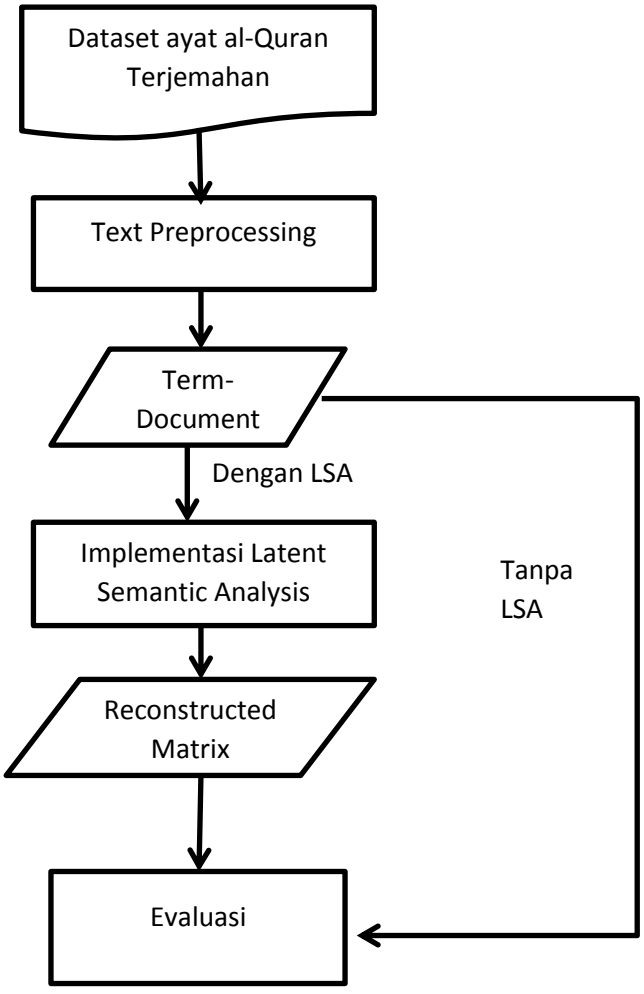
$$\text{Precision} = \frac{tp}{tp + fp} \quad (4)$$

Dimana :

tp : *True Positive*

tn : *True Negative*

## 2.4. Perancangan Sistem




**Gambar 1 Alur Proses Sistem**

Proses pertama akan mengolah ayat al-Quran untuk dilakukan *preprocess* dan hasilnya akan digunakan untuk pengolahan matriks dengan *singular value decomposition* untuk kemudian direkonstruksi menjadi matriks baru. Matriks ini yang akan dibandingkan dengan matriks awal sebelum pemrosesan *latent semantic analysis*, apakah matriks hasil pemrosesan menghasilkan perhitungan nilai *similarity* lebih baik dari ayat-ayat yang secara natural bisa diketahui perbedaan maknanya.

### 3. Pembahasan

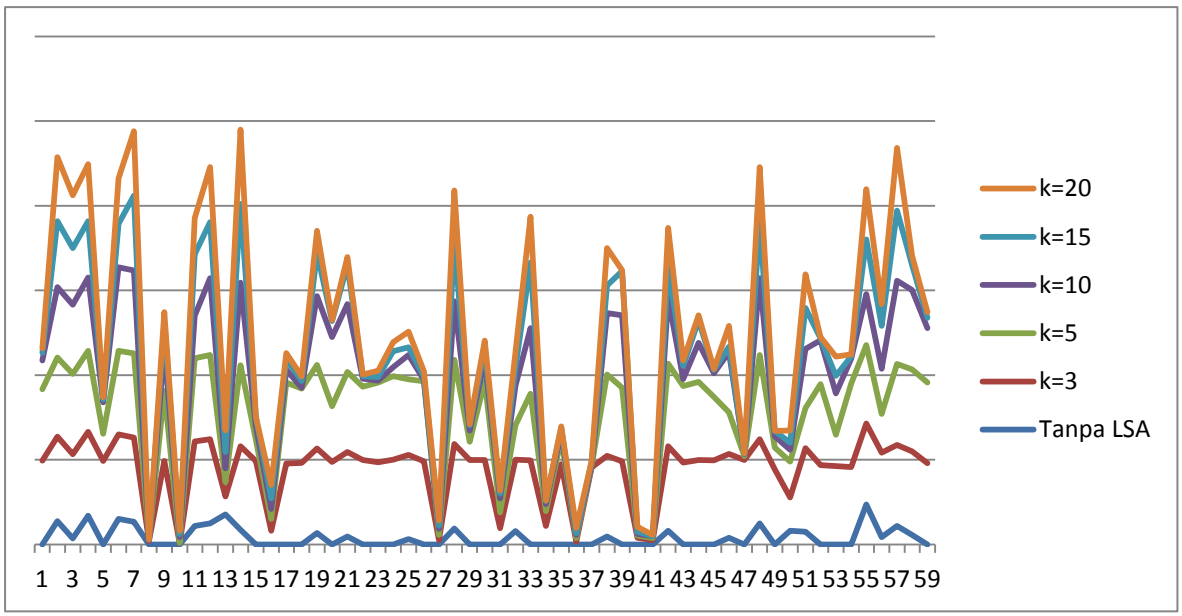
Dari pengujian yang dilakukan, dimana melibatkan 59 sample dan 6 nilai dimensi, didapat hasil sebagai berikut,

**Tabel Rekapitulasi Hasil Pengujian**

| | Akurasi | F-Measure (efektifitas hasil) |
|---|---|---|
| Tanpa LSA | 85% | 0% |
| | | |
| LSA K=3 | 41% | 32% |
| | | |
| LSA K=5 | 44% | 27% |
| | | |
| LSA K=10 | 64% | 40% |
| | | |
| LSA K=15 | 68% | 24% |
| | | |
| LSA K=20 | 71% | 19% |

Hasil diatas menunjukkan bahwa meskipun akurasi sistem yang tidak menggunakan LSA lebih besar, tetapi kualitasnya sangat buruk yaitu 0%. Hal ini terjadi akibat kecendrungan nilai similarity yang selalu kecil. Sehingga menyebabkan semua pasangan ayat yang diuji diklasifikasikan sebagai ayat yang maknanya berbeda, sedangkan jumlah ayat yang menurut *goldstandard* maknanya berbeda memamng lebih banyak. Inilah yang menyebabkan akurasinya tinggi tapi tidak bernilai apapun karena tidak bisa mendeteksi sesuatu yang ingin kita cari yaitu ayat yang memiliki kesamaan makna. Di sisi lain, penggunaan LSA meningkatkan F-measure hingga 40%, yaitu pada LSA dengan K=10. Hal ini terbilang baik, mengingat LSA adalah *unsupervised learning* yang tidak menggunakan *knowledge* tambahan untuk menentukan kesamaan.

**Grafik Nilai Similarity dari Tiap Dimensi**




Dari grafik diatas terlihat bahwa dimulai dari K=5 dan seterusnya, grafiknya tidak terlalu berbeda jauh. Artinya persebarannya mulai stabil dan bisa digunakan. Semakin kecil dimensinya maka nilai *similarity* semakin seragam dan tidak mencerminkan isi konsep yang sesungguhnya.

## 4. Kesimpulan

Dari sekumpulan pengujian yang dilakukan, secara keseluruhan penerapan *latent semantic analysis* sudah cukup efektif untuk mencari kesamaan antar ayat pada al-Quran yang berupa *short text*. Terlihat dari nilai F-Measure yang paling tinggi adalah 14%. Hal ini signifikan dibandingkan dengan metode konvensional yaitu *term-document* matriks tanpa LSA yang F-measurenya sampai 0%. Selain itu, hasil ini dianggap cukup baik karena LSA tidak menggunakan knowledge tambahan seperti Wordnet dan thesaurus dan sehingga hasilnya tidak terlalu tinggi dan tapi tidak terlalu rendah pula.

Selain itu, penentuan dimensi juga sangat berpengaruh terhadap kualitas sistem. Dapat disimpulkan bahwa semakin kecil dimensi yang digunakan, maka nilainya akan semakin seragam dan tidah relevan dgn data asli. Selain itu, ada titik stabil, dimana mulai dimensi tersebut maka dimensi yang lebih besar tidak akan terlalu jauh perubahannya.

# DAFTAR PUSTAKA


[1] Preslav Nakov , "Latent Semantic Analysis of Textual Data".

[2] Thomas K Landauer, Peter W Foltz, and Darrell Laham, "An Introduction to Latent Semantic Analysis," *Disscourse Processes*, vol. 25, pp. 259-284, 1998.

[3] Aminul Islam and Inkpen Diana, "Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity," *ACM Transactions on Knowledge Discovery from Data*, p. 5, 2008.

[4] Dudi Hermawandi, IMPLEMENTASI PEMBOBOTAN PADA ESSAY GRADING METODE LATENT SEMANTIC ANALYSIS SEBAGAI PENUNJANG E-LEARNING, 2008.

[5] Dharmendra Kanejiya, Arun Kumar, and Surendra Prasad, "Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA".

[6] Thomas K Landauer and Susan T Dumais, "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation o f Knowledge," *American Psychological Association*, vol. 104, no. 2, pp. 211-240, 1997.

[7] Bob Rehder and Darrell Laham, "Latent Semantic Analysis And Knowledge Assesment".

[8] Gerald Salton, A Wong, and C S Yang, "A vector space model for automatic indexing," *Communications of ACM*, vol. 18, no. 11, 1975.

[9] Erica Chisholm and Tamara G Kolda, "New Term Weighting Formulas for the Vector Space Method in Information Retrieval," *Oak Ridge National Laboratory*, 1999.

[10] Sabrina Simmons and Zachary Estes, Using latent semantic analysis to estimate similarity.

[11] Ah Hwee Tan, Text Mining : The state of the art and the challenges, 1999.