IMPLEMENTASI *EXPLICIT SEMANTIC ANALYSIS* BERBAHASA INDONESIA MENGGUNAKAN *CORPUS* WIKIPEDIA INDONESIA

Faisal Rahutomo[1], Pramana Yoga Saputra[2], Carfin Febriawan Pratama Putra [3]

Jurusan Teknologi Informasi, Program Studi Teknik Informatika, Politeknik Negeri Malang
faisal@polinema.ac.id[1], pramanay@gmail.ac.id[2], cfebriawan@hotmail.com[3]

Abstrak

Pengembangan terhadap Ujian Online Bahasa Indonesia dalam bentuk esai masih terus dilakukan sampai sekarang guna memperoleh nilai akurasi yang lebih baik dalam memberikan suatu penilaian. Penilaian yang sudah ada saat ini masih menggunakan kemiripan kata pada teks kunci jawaban dan teks jawaban. Cara tersebut memiliki kelemahan mengingat kata dengan tulisan berbeda dapat memiliki makna yang sama. Masalah tersebut dapat diatasi menggunakan skema vektor konsep. Vektor konsep bekerja pada level makna dari sebuah kata. Skema vektor konsep ini dapat diimplementasikan salah satunya menggunakan metode *Explicit Semantic Analysis* (ESA). Metode ESA memerlukan sebuah korpus yang besar, penelitian ini akan menggunakan korpus dari Artikel Wikipedia Indonesia. Dengan menggunakan metode ESA proses penilaian akan dilakukan dengan membandingkan kemiripan makna dari teks kunci jawaban dengan teks jawaban. Pengujian dilakukan dengan membandingkan 400 teks jawaban soal esai *online* dengan kunci jawabannya. Dari hasil pengujian tersebut didapatkan kesimpulan bahwa nilai *percentage error* metode ESA adalah 65%, di mana angka tersebut merupakan probabilitas *error* yang terlalu tinggi. Pengujian lain yang dilakukan adalah dengan membandingkan nilai *percentage error* metode ESA dengan metode lain seperti *Cosine Similarity*, *Euclidean Distance*, dan *Jaccard* yang memberikan konklusi bahwa metode ESA tidaklah lebih akurat dari metode-metode lain tersebut.

Kata kunci : *explicit semantic analysis*, pemrosesan teks, soal esay online, wikipedia indonesia

1. Pendahuluan

Setiap proses pembelajaran membutuhkan suatu proses evaluasi sebagai sarana untuk menentukan tingkat pemahaman siswa. Evaluasi yang dapat diterapkan cukup beragam seperti menggunakan soal pilihan ganda, isian singkat hingga esai. Beberapa penelitian mengungkapkan bahwa soal pilihan ganda dan isian singkat masih dirasa kurang memadai dalam mengukur tingkat pemahaman siswa. Sebaliknya, evaluasi materi pembelajaran dalam bentuk esai akan lebih akurat dalam menentukan tingkat pemahaman siswa karena siswa diharuskan menuliskan informasi-informasi yang diketahui secara detail dalam menjawab pertanyaan yang diajukan.

Saat ini, perkembangan sistem informasi untuk membantu dalam mengevaluasi pemahaman siswa terhadap proses belajar mengajar sudah cukup banyak. Sistem informasi tersebut telah memiliki fasilitas ujian *online* baik dalam bentuk pilihan ganda, jawaban singkat, dan esai. Namun, pengembangan terhadap ujian *online* dalam bentuk esai masih terus dilakukan sampai sekarang. Pengembangan tersebut perlu dilakukan agar didapatkan nilai akurasi yang lebih baik dalam memberikan suatu penilaian.

Telah banyak penelitian yang telah dilakukan untuk menemukan metode yang paling baik dalam menghitung penilaian hasil jawaban soal esai *online*. Beberapa penelitian mengenai ujian esai berbahasa Indonesia adalah sebagai berikut Penilaian Esai Jawaban Bahasa Indonesia Menggunakan Metode SVM-LSA dengan Fitur Generik yang menghasilkan keakurasian 73% (Adhitia, 2009), *An Attempt to Create an Automatic Scoring Tool for Short Text Answer in Bahasa Indonesia* pada aplikasi moodle memiliki standar deviasi sebesar 3-30 dari berbagai macam jenis data yang diujikan (Thamrin, 2004), Perbandingan Algoritma TF/IDF dan BLEU untuk Penilaian Jawaban Esai Otomatis yang menghasilkan nilai korelasi 0.7 (Nugroho, 2014), Analisis Aspek-Aspek Ujian Esai Daring Berbahasa Indonesia yang telah menggunakan perbandingan antara vektor-vektor kemiripan antar kedua teks (Trisna, 2016).Dari beberapa hasil penelitian tersebut dapat dilihat bahwa penelitian menggunakan skemavektor konseppada teks masih jarang digunakan. Vektor konsep bekerja pada level makna dari sebuah kata, artinya dua kata yang memiliki tulisan berbeda bisa saja memiliki makna yang sama (misal: kata yang bersinonim). Dalam kasus ini adalah kemiripan antara teks kunci jawaban soal esai dengan teks hasil jawaban siswa.

Metode yang digunakan untuk menerapkan skema vektor konsep adalah *Explicit Semantic Analysis* (ESA). Metode ESA memerlukan sebuah matriks yang berisikan vektor kata. Vektor kataini diperoleh melalui *text pre-processing* dari sebuah *corpus*teks yang sangat besar yang dalam hal ini penulis menggunakan artikel Wikipedia bahasa Indonesia. Penulis melakukan penelitian tentang "Implementasi Explicit Semantic Analysis (ESA) berbahasa Indonesia menggunakan corpus Wikipedia Bahasa Indonesia" guna meningkatkan tingkat akurasi penilaian jawaban soal esai online.

2. Tinjauan Pustaka

2.1 *Text Mining*

Text mining adalah proses menganalisis teks untuk mengekstrak informasi yang berguna untuk tujuan tertentu. *Text mining* memiliki tugas yang lebih kompleks karena melibatkan data teks yang sifatnya tidak terstruktur dan kabur (*fuzzy*). *Text mining* merupakan bidang multi-disiplin yang melibatkan analisis teks, ekstraksi informasi, *clustering*, kategorisasi, visualisasi, teknologi basis data, *machine learning*, dan *data mining*. Perbedaan mendasar antara *text mining* dan *data mining* terletak pada sumber data yang digunakan. Pada *data mining*, pola-pola diekstrak dari basis data yang terstruktur, sedangkan di *text mining*, pola-pola diekstrak dari data tekstual (*natural language*).

1. *Case Folding* dan *Tokenizing. Case folding* adalah mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf **a** sampai dengan **z** yang diterima. Karakter selain huruf dihilangkan dan dianggap *delimiter*. Tahap *tokenizing* adalah tahap pemotongan *string input* berdasarkan tiap kata yang menyusunnya.
2. *Filtering. Filtering* adalah tahap mengambil kata-kata penting dari hasil proses *tokenizing*. Terdapat beberapa algoritma dalam *filtering* yaitu *stop-list* dan *word-list*. Algoritma *stop-list* merupakan algoritma yang digunakan untuk mengeliminasi kata-kata yang tidak deskriptif. Algoritma *word-list* adalah algoritma yang digunakan untuk menyimpan kata-kata memiliki nilai deskriptif.
3. *Stemming. Stemming* adalah proses untuk memecahkan setiap varian-varian suatu kata menjadi kata dasar. Proses *stemming* pada kata Bahasa Indonesia berbeda dengan *stemming* pada kata Bahasa Inggris. Proses *stemming* pada kata Bahasa Inggris adalah proses untuk mengeliminasi sufiks pada kata sementara proses *stemming* pada Bahasa Indonesia adalah proses untuk mengeliminasi sufiks, prefiks dan konfiks. Terdapat beberapa algoritma dalam *stemming* bahasa Indonesia, antara lain algoritma Porter dan algoritma Nazief & Adriani.
4. *Analyzing.* Tahap *Analyzing* merupakan tahap penentuan seberapa jauh kemiripan antar dokumen teks. Terdapat beberapa metode untuk menentukan kemiripan antar dokumen teks antara lain metode *Euclidean Distance*, metode *Cosine Similarity*, metode *Jaccard Coefficient*, dll. Metode tersebut menggunakan persamaan matematika dalam menentukan nilai kemiripan antar berkas dokumen teks.

2.2 *Explicit Semantic Analysis* (ESA)

Metode ESA merupakan metode yang merepresentasikan sebuah *world knowledge* yang besar ke dalam sebuah matriks kata (*term*). Dalam hal ini, *world knowledge* yang digunakan adalah artikel Wikipedia bahasa Indonesia. Matriks yang digunakan adalah matriks **I** dengan ukuran ***n x m***, di mana ***n*** adalah jumlah kata dan ***m*** adalah jumlah artikel atau konsep yang ada pada Wikipedia bahasa Indonesia.

$$\mathbf{I} = \begin{pmatrix} i_{1,1} & \cdots & i_{1,2} & \cdots & i_{1,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ i_{2,1} & \cdots & i_{2,2} & \cdots & i_{2,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ i_{n,1} & \cdots & i_{n,2} & \cdots & i_{n,m} \end{pmatrix} \quad (1)$$

Kemudian, metode ESA akan men-*transpose* matriks tersebut menjadi matriks $\mathbf{I}^{\mathrm{T}}$ dengan ukuran ***m x n***.

$$\mathbf{I}^{\mathrm{T}} = \begin{pmatrix} i_{1,1} & \cdots & i_{1,2} & \cdots & i_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ i_{2,1} & \cdots & i_{2,2} & \cdots & i_{2,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ i_{m,1} & \cdots & i_{m,2} & \cdots & i_{m,n} \end{pmatrix} \quad (2)$$

Setelah matriks $\mathbf{I}^{\mathrm{T}}$ terbentuk, terdapat dua tahap yang harus dilakukan, yang pertama yaitu melakukan konversi dari vektor *term***x** (dengan panjang ***n***) menjadi vektor konsep **v** (dengan panjang ***m***) dengan cara melakukan proses perkalian antara matriks $\mathbf{I}^{\mathrm{T}}$ dengan vektor **x**.

$$\mathbf{v} = \mathbf{I}^{\mathrm{T}}\,\mathbf{x} \quad (3)$$

Yang dapat digambarkan secara detail sebagai berikut:

$$\begin{pmatrix} v_1 \\ \vdots \\ v_2 \\ \vdots \\ v_m \end{pmatrix} = \begin{pmatrix} i_{1,1} & \cdots & i_{1,2} & \cdots & i_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ i_{2,1} & \cdots & i_{2,2} & \cdots & i_{2,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ i_{m,1} & \cdots & i_{m,2} & \cdots & i_{m,n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad (4)$$

Setelah itu dilanjutkan ke tahap kedua yaitu dilakukan tahap pengukuran persamaan antara dua buah vektor konsep **u** dan **v** yang dilakukan menggunakan rumus *cosine similarity*.

2.3 *Cosine Similarity*

Cosine Similarity merupakan metode perhitungan jarak antara vektor A dan B yang menghasilkan sudut *cosine x* di antara kedua vektor tersebut. Nilai sudut kosinus antara dua vektor menentukan kesamaan dua buah objek yang dibandingkan di mana nilai terkecil adalah 0 dan nilai terbesar adalah 1. Berikut rumus metode perhitungan *Cosine Similarity*:

$$\cos(\theta) = \frac{A \cdot B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2}\sqrt{\sum_{i=1}^{n} {B_i}^2}} \quad (5)$$

Dengan $A \cdot B$ merupakan *dot product*. *Dot product* merupakan nilai yang mengekspresikan sudut antara dua vektor. *Dot product* merupakan skalar nilai hasil dari operasi dua vektor yang memiliki jumlah komponen yang sama. Jika vektor A dan B memiliki komponen sebanyak *n*, maka *dot product* dapat dihitung dengan rumus berikut:

$$A \cdot B = A_1 B_1 + \cdots + A_n B_n \quad (6)$$

Dot product dapat dihitung dengan menjumlahkan *product* dari masing-masing komponen pada kedua vektor. Jika vektor A dan vektor B merupakan vektor 3 dimensi, maka perhitungan *dot product* adalah sebagai berikut:

$$A \cdot B = A_x * B_x + A_y * B_y + A_z * B_z \quad (7)$$

Sedangkan $|A|$ merupakan panjang vektor. Panjang vektor dapat dihitung dengan rumus sebagai berikut :

$$|A| = \sqrt{{X_1}^2 + {X_2}^2 + {X_3}^2} \quad (8)$$

Perhitungan untuk menentukan nilai persentase kemiripan antar dokumen teks, maka persentase kemiripan didapat dengan mengalikan nilai *Cosine Similarity* terhadap 100. Berikut rumus untuk menentukan nilai persentase kemiripan:

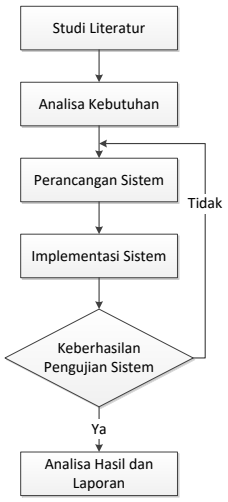
$$Kemiripan(\%) = cos(A, B) \cdot 100 \quad (9)$$

3. Metodologi

Metode penelitian yang digunakan dalam merancang Sistem Aplikasi untuk Implementasi *Explicit Semantic Analysis* (ESA) Berbahasa Indonesia Menggunakan *Corpus* Wikipedia Indonesia, yang dapat dilihat pada Gambar 1.

3.1 Studi Literatur

Pada tahapan ini penelitian dilakukan dengan cara mempelajari berbagai literatur melalui pengumpulan dokumen-dokumen, referensi-referensi, buku-buku, sumber dari internet, maupun sumber lain yang mendukung dan diperlukan dalam perancangan sistem.



Gambar 1. Metodologi pelaksanaan

3.2 Analisa Kebutuhan

Terdapat beberapa analisa kebutuhan antara lain meliputi kebutuhan perangkat lunak (*software*), perangkat keras (*hardware*), kebutuhan pengguna sistem (*brainware*), dan *file dump* artikel Wikipedia bahasa Indonesia. Spesifikasi minimum perangkat lunak yang digunakan dalam penelitian ini dapat dilihat pada Tabel 1. Sedangkan spesifikasi minimum perangkat keras yang digunakan padapenelitian ini dapat dilihat pada Tabel 2.

Tabel 1. Spesifikasi minimum perangkat lunak

Perangkat Lunak	**Keterangan**
Windows 7	Sistem Operasi yang digunakan untuk menjalankan sistem.
Sublime3 atau PhpStorm	IDE / Aplikasi teks editor untuk menulis kode program
Apache HTTP Server	Sebagai Web Server dari *website* yang akan dibuat.
MySQL	Sebagai DBMS yang akan digunakan.
PHP	Sebagai bahasa pemrograman utama yang digunakan.
Python	Sebagai bahasa pemrograman untuk proses ekstraksi *file dump* Wikipedia.

Tabel 2. Spesifikasi minimum perangkat keras

Perangkat Keras	**Keterangan**
Processor	Intel Core i3 2.0 GHz
RAM	4GB
Hard disk	10GB
Monitor	Disesuaikan
Perangkat *Input*	*Mouse*, *Keyboard*

Kebutuhan Pengguna Sistem (*Brainware*) adalah, pengguna berinteraksi secara langsung dengan sistem. Berdasarkan tugas dan fungsinya dalam sistem ini dibagi menjadi 2 (dua) pengguna sistem. Daftar pengguna sistem dapat dilihat pada Tabel 3.

Tabel 3. Pengguna sistem

No.	Pengguna	Hak Akses
1.	Administrator	Memiliki akses terhadap semua fasilitas sistem.
2.	Pengguna Umum	Masuk dan keluar sistem, melakukan perbandingan dua teks.

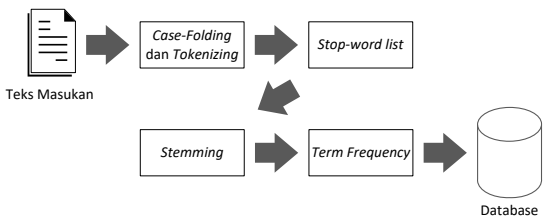
Dalam penelitian ini dibutuhkan sebuah *worldknowledge* yang cukup besar yaitu artikel-artikel yang ada dalam Wikipedia bahasa Indonesia. Wikipedia sudah menyediakan sebuah fasilitas untuk mengunduh data artikel dalam bentuk XML atau format lainnya. Artikel yang akan digunakan adalah artikel Wikipedia bahasa Indonesia tanggal 20 Desember 2016. *File dump* tersebut dapat diunduh melalui *link*https://dumps.wikimedia.org/idwiki/20161220/.Setelah merujuk *link* tersebut, maka dapat dipilih *file* XML yang berisikan seluruh artikel terbaru, yaitu dengan memilih *link*https://dumps.wikimedia.org/idwiki/20161220/idwiki-20161220-pages-articles-multistream.xml.bz2.

3.3 Perancangan Sistem

Dalam penelitian ini, penulis membagi menjadi beberapa tahapan proses perencanaan. Berikut adalah tahapan-tahapan yang digunakan.

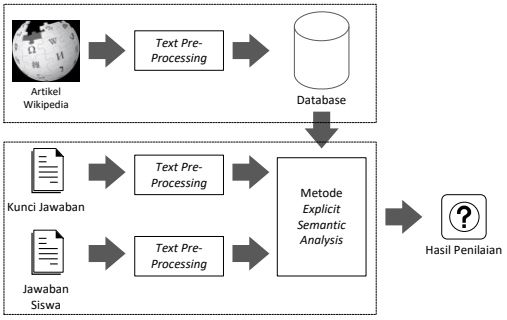
3.3.1. *Text Pre-Processing*

Pada tahapan ini teks artikel Wikipedia maupun teks data uji perlu dilakukan sebuah ekstraksi teks sehingga berubah menjadi data numerik yang dapat diolah yang dapat dilihat pada Gambar 2.



Gambar 2. Proses *Text Pre-Processing*

Pada Gambar 2, teks masukan yang digunakan adalah artikel Wikipedia bahasa Indonesia dan juga teks data uji (kunci jawaban dan jawaban siswa). Proses *Text Pre-Processing* pada sistem dibagi menjadi dua kelompok besar berdasarkan alur waktu penerapannya. Dua kelompok proses tersebut dapat dilihat pada Gambar 3.

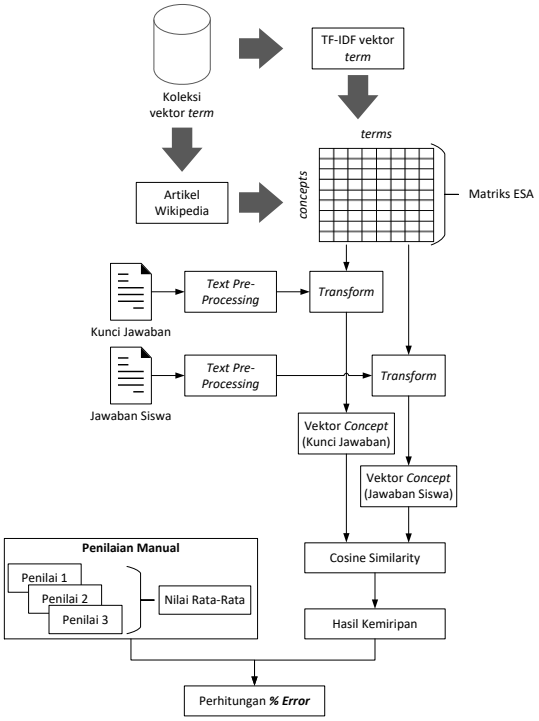


Gambar 3. Pembagian proses *Text Pre-processing*

Pada Gambar 3 dapat dilihat bahwa ketiga jenis dokumen teks masukan yaitu artikel Wikipedia Indonesia, kunci jawaban, dan jawaban siswa diproses terlebih dahulu melalui tahapan *text pre-processing*. Namun, tahapan *text pre-processing* untuk ketiga jenis masukan tersebut dilakukan pada dua kelompok waktu yang berbeda. Untuk kelompok pertama yaitu tahapan *text pre-processing* pada artikel Wikipedia Indonesia dilakukan hanya satu kali proses saja. Pada kelompok pertama ini tahapan *text pre-processing* dilakukan untuk membuat indeks matriks yang dibutuhkan pada saat menerapkan metode ESA. Setelah indeks matriks terbentuk, maka proses *text pre-processing* pada kelompok proses pertama ini tidak akan dilakukan kembali, proses yang dilakukan adalah proses pada kelompok kedua yaitu proses *text pre-processing* untuk kunci jawaban dan jawaban siswa. Proses pada kelompok kedua ini dilakukan dapat dilakukan berulang kali.

3.3.2. Implementasi ESA

Setelah data-data vektor kata sudah disimpan ke dalam *database*, maka dapat dilakukan pengimplementasian metode ESA terhadap vektor-vektor kata tersebut yang dapat dilihat pada Gambar 4.Langkah pertama yaitu mengambil koleksi vektor kata Wikipedia dari *database* dan membentuknya menjadi indeks matriks di mana nilai TF-IDF vektor katamenjadi nilai kolom dari matriks dan artikel-artikel Wikipedia menjadi nilai baris dari matriks. Indeks matriks yang terbentuk kemudian akan dikalikan dengan vektor kata teks data uji dalam hal ini adalah vektorkatakunci jawaban dan vektor katajawaban siswa. Hasil perkalian kedua vektor tersebut akan menghasilkan dua vektor konsepyaitu vektor konsep kunci jawaban dan vektor konsep jawaban siswa. Kedua vektor konsep inilah yang nantinya digunakan sebagai parameter dalam menghitung kemiripan kedua teks yang dibandingkan.



Gambar 4. Rancangan proses ESA

3.3.3. Pengujian Kemiripan Teks

Setelah didapatkannya dua buah vektor konsep, maka dilakukan proses perhitungan kemiripan makna antara dua vektor tersebut, yang mana dalam penelitian ini menggunakan rumus *cosine simiarity*. Hasil nilai kemiripan tersebut kemudian dibandingkan dengan hasil penilaian manual. Perbandingan tersebut akan menghasilkan nilai persentase *error* terhadap pengujian sistem. Semakin kecil nilai *error* yang didapatkan artinya penilaian yang dilakukan oleh sistem telah sesuai dengan penilaian yang dilakukan secara manual.

4. Implementasi

4.1 Ekstraksi *File Dump* Wikipedia

*File dump*Wikipedia yang digunakan dalam penelitian ini adalah *file dump* yang berisikan artikel.Teks artikel tersebut masih belum dalam bentuk *plaintext* yang sudah siap diproses melainkan dalam bentuk teks yang mengandung kode-kode tertentu yang disebut kode *template* wikipedia.Untuk menghilangkan kode *template* tersebut digunakan sebuah ekstraktor. Ekstraktor yang digunakan adalah WikiExtractor versi 2.75 yang merupakan aplikasi *third party* yang dapat diunduh melalui GitHub.

Hasil ekstraksi WikiExtractor versi 2.75 merupakan sebuah file XML yang berisikan teks artikel yang sudah bersih dari kode template wikipedia dan berukuran jauh lebih kecil dari *file dump* wikipedia.

4.2 Memasukkan Hasil Ekstraksi ke *Database*

Hasil ekstraksi artikel kemudian dimasukkan ke dalam *database* melalui sistem yang dibuat.Proses ini membutuhkan waktu sekitar 11 menit dengan total artikel sejumlah 372.829 artikel.

4.3 Ekstraksi Kata

Setelah data artikel berhasil dimasukkan, kemudian dilakukan proses ekstraksi kata dari keseluruhan data artikel. Proses ini membutuhkan waktu sekitar 9 jam dengan total kata sejumlah 733.281 kata.

4.4 Menghitung Nilai TF*IDF

Setelah data artikel dan kata siap, kemudian dilakukan proses perhitungan nilai TF*IDF. Proses perhitungan nilai TF*IDF dilakukan dalam 3 tahap, yaitu tahap perhitungan nilai TF, perhitungan nilai IDF, dan terakhir perhitungan nilai TF*IDF. Nilai TF*IDF inilah yang menjadi nilai pada indeks matriks ESA.

4.5 Menghitung Kemiripan Makna

Setelah indeks matriks berhasil dibuat maka dapat dilakukan perhitungan nilai kemiripan makna dua buah teks. Perhitungan dilakukan dengan cara memasukkan dua buah teks yang ingin dicari kemiripan maknanya. Proses ini membutuhkan waktu sekitar 2-5menit. Proses perhitungan cukup memakan waktu karena dipengaruhi oleh keterkaitan kedua teks yang diuji dengan data pada indeks matriks.

5. Pengujian

Pengujian akurasi sistem bertujuan untuk menghitung tingkat keakurasian dari hasil perhitungan oleh sistem. Perhitungan keakurasian sistem dilakukan dengan menggunakan perhitungan *percentage error*, di mana dilakukan perhitungan perbandingan antara penilaian manual oleh manusia dengan penilaian sistem. Pada penelitian ini untuk menguji tingkat akurasi sistem digunakan data uji penelitian sebelumnya yaitu data uji ujian esai *online*. Di mana terdapat 400 teks jawaban siswa yang sudah dinilai secara manual oleh seorang guru dan juga sudah dinilai secara sistem menggunakan metode-metode pada penelitian sebelumnya.

Pada penelitian ini pengujian akurasi dilakukan dengan cara menghitung nilai kemiripan teks kunci jawaban dan teks jawaban siswa. Setalah itu dicari nilai *percentage error*, semakin kecil nilai *percentage error* yang dihasilkan maka semakin akurat penilaian yang dilakukan oleh sistem. Pengujian lain yang dilakukan adalah membandingkan nilai rata-rata *percentage error*

metode ESA dengan metode-metode lain pada penelitian sebelumnya. Tujuan dari pembandingan tersebut adalah untuk mengetahui apakah metode ESA terbukti lebih akurat dari metode-metode sebelumnya. Pada penelitian sebelumnya metode yang digunakan adalah *Cosine Similarity, Euclidean Distance, Jaccard, Cosine Similaruty dengan Stemming, Euclidean Distance dengan Stemming,* dan *Jaccard dengan Stemming*.

Tabel 4. Hasil Analisa

Metode	Lifestyle	Olahraga	Politik	Teknologi	Rata-rata
% Err Cos	77,4	32,8	38,1	45,5	48,4
% Err Euc	253,9	81,1	82,6	134,3	138,0
% Err Jac	66,1	51,7	48,5	53,3	54,9
% Err Cos Stemm	75,0	32,3	40,0	42,9	47,6
% Err Euc Stemm	252,1	80,6	81,1	131,6	136,4
% Err Jacc Stemm	61,8	48,6	45,8	51,0	51,8
% Err ESA	96,7	50,3	58,8	56,6	65,6

Dari keempat kategori pertanyaan, hasil nilai *percentage error* rata-rata keseluruhan dapat dilihat pada tabel 4 di atas. Dari tabel 4 tersebut dapat dilihat bahwa nilai *percentage error* metode ESA adalah 65% dan menempatkan metode ESA pada peringkat kelima, itu artinya bahwa hipotesis penulis yang beranggapan bahwa dengan memanfaat metode ESA yang mampu bekerja pada level makna kata tidak terbukti dapat meningkatkan nilai akurasi dalam menghitung penilaian soal esai *online* atau secara umum akurasi perhitungan kemiripan makna.

6. Kesimpulan dan Saran

Berdasarkan analisa dan perancangan, implementasi serta pengujian sistem, maka didapatkan kesimpulan sebagai berikut.

Dengan menerapkan *text-preprocessing* pada teks artikel Wikipedia maka dapat dihasilkan sebuah indeks matriks ESA yang bisa dimanfaat untuk penelitian-penelitian yang berhubungan dengan analisa *semantic* salah satu contohnya adalah perhitungan kemiripan makna.

Penggunaan metode ESA untuk meningkatkan akurasi penilaian ujian esai *online* menghasilkan sebuah konklusi yang bertolak belakang dengan hipotesis yang diharapkan penulis, karena setelah dibandingkan dengan hasil perhitungan metode-metode sebelumnya nilai *percentage error* metode ESA dianggap masih cukup besar yaitu 65% dan tidak lebih baik atau akurat dengan metode-metode sebelumnya.

Pada penelitian ini ada beberapa saran yang dapat diberikan untuk pengembangan sistem selanjutnya meliputi beberapa hal sebagai berikut.

Penggunaan WikiExtractor untuk membersihkan teks artikel Wikipedia masih dirasa kurang baik karena ada beberapa bagian penting dari teks artikel Wikipedia yang hilang salah satunya adalah teks pada bagian InfoBox.

Hasil ekstraksi kata pada penelitian ini berjumlah lebih dari 700.000 sekian data kata, yang di mana tidak semua kata tersebut merupakan kata yang valid menurut EYD maupun tidak termasuk kata dari suatu istilah. Sehingga bisa dikatakan mayoritas kata yang didapatkan dari hasil ekstraksi merupakan kata yang tidak valid. Diharapkan pada pengembang selanjutnya sistem yang dibuat mampu memberi *tagging* untuk mengetahui mana kata yang valid dan tidak valid.

Daftar Pustaka:

Adhitia, R. & Purwarianti, A. *"Penilaian Esai Jawaban Bahasa Indonesia Menggunakan Metode SVM-LSA dengan Fitur Generik"*. Journal of Information Systems, Volume 5, Issues 1, pp 33. 2009

Azhar Firdaus, Ernawati, dan Arie Vatresia. *"Aplikasi Pendeteksi Kemiripan Pada Dokumen Teks Menggunakan Algoritma Nazief & Adriani Dan Metode Cosine Similarity"*. Jurnal Teknologi Informasi, Volume 10 Nomor 1, April 2014

Nugroho, H. W., dkk. *"Perbandingan Algoritma TF/IDF dan BLEU untuk Penilaian Jawaban Esai Otomatis"*. Edu Komputika Jurnal, pp 43. 2014

Rahutomo, Faisal, et al. *"Econo-ESA reduction scheme and the impact of its index matrix density"*. Procedia Computer Science 35, 2014

Roshinta, Trisna Ari. *"Analisa Aspek-Aspek Ujian Esai Daring Berbahasa Indonesia"*. Seminar Nasional Terapan Riset Inovatif Semarang, Volume 1. 2016

Thamrin, H., Wantoro, J. *"An Attempt to Create an Automatic Scoring Tool of Short Text Answer in Bahasa Indonesia"*. Proceeding of Internasional Conference on Electrical Engineering, Computer Science and Informatics, pp 96-98. 2014

Witten, Ian H. 2001. *"Adaptive Text Mining: Inferring Structure from Sequences"* [Online]. Tersedia: http://www.cs.waikato.ac.nz/~ihw/papers/01IHW-Adaptivetextmining.pdf