JURNAL SWABUMI, Vol.12 No.1 Maret 2024, pp. 39~46
ISSN: 2355-990X
E-ISSN: 2549-5178

Penerapan Information Retrieval dalam Sistem Analisis Kemiripan Proposal Skripsi menggunakan Cosine Similarity

Muhammad Dzikry Afandi[1], Ahmad Homaidi[2], Abd Ghofur[3], Ach Zubairi[4]

[1,2,3,4] Universitas Ibrahimy
e-mail: [1]afandifa24@gmail.com [2]ahmadhomaidi@gmail.com [3]apunkbwi@gmail.com
[4]dsnfsei01@gmail.com

Abstrak

Di lingkungan akademik, konsep plagiarisme cukup sering dibahas, plagiarisme yang biasa dilakukan adalah plagiarisme terhadap karya tulis ilmiah. Masalah yang sama juga dihadapi oleh Universitas Ibrahimy Sukorejo khususnya Fakultas Sains dan Teknologi. Mahasiswa biasanya hanya mengubah tempat penelitian, hal ini tentunya mengkhawatirkan sebab tidak ada kebaharuan yang ditawarkan pada penelitian tersebut. Penelitian ini menawarkan sebuah sistem yang bisa menganalisa sejauh mana kemiripan dari proposal skripsi yang diajukan oleh mahasiswa. Sistem menerima masukan berupa judul dan ringkasan proposal skripsi yang akan diajukan. Sistem temu kembali atau information retrieval akan digunakan bersamaan dengan text mining. Penelitian ini meggunakan metode perangkingan cosine similarity yang sebelumnya dilakukan pembobotan term frequency-inverse document frequency. Hasilnya sistem mampu untuk mengenali masukan yang diberikan pengguna dan memberikan perangkingan sesuai dengan perhitungan yang dihasilkan. Pada pengujian sistem mendapatkan kemiripan sebesar 100% untuk query yang sama persis dengan dataset. Sedangkan pada query yang di dimasukkan agak mirip, sistem memberikan persentase sebesar 18.4%. Sistem juga bisa memberikan gambaran kepada panitia skripsi tentang penelitian yang akan dilakukan oleh mahasiswa.

Kata Kunci: cosine similarity, sistem temu kembali, proposal skripsi, text mining, tf-idf.

Abstract

In academic circles, the concept of plagiarism is often discussed, the plagiarism that is usually done is plagiarism against scientific papers. The same problem is also faced by Ibrahimy Sukorejo University, especially the Faculty of Science and Technology. Students usually only change the place of research, this is of course worrying because there is no innovation offered in the research. This research offers a system that can analyze how far the thesis proposals submitted by students are similar. The system receives input in the form of a title and a summary of the thesis proposal to be submitted. Retrieval system or information retrieval will be used in conjunction with text mining. This study uses the cosine similarity ranking method, which was previously weighted by term frequency-inverse document frequency. As a result, the system is able to recognize the wishes given by the user and provide a ranking according to the resulting calculations. In testing, the system obtained a similarity of 100% for queries that were exactly the same as the dataset. Meanwhile, the query entered is somewhat similar, the system gives a percentage of 18.4%. The system can also provide an overview to the thesis committee about the research that will be carried out by students.

Keywords*: cosine similarity, information retrieval, thesis proposal, text mining, tf-idf.*

1. Pendahuluan

Skripsi biasanya dikaitkan dengan sebuah karya tulis ilmiah yang dibuat oleh mahasiswa strata satu sebagai hasil penelitian yang ditujukan untuk melatih sikap dan pemecahan masalah dengan cara mengaplikasikan ilmu atau teori yang diperolehnya untuk memecahkan masalah. (Cahyani & Arif, 2022; Hadi, 2016).

Diterima Agustus 16, 2023; Revisi September 17 , 2023; Disetujui November 19, 2023

Mahasiswa baru bisa mengerjakan proposal skripsi setelah menyelesaikan 120 sks/7 semester, dengan maksimal keberadaan nilai C pada tiga mata kuliah, dan telah mendapatkan persetujuan dari calon dosen pembimbing. umumnya ada tiga bab yang dicakup oleh proposal penelitian. Bab pendahuluan, bab kajian teori atau tinjuan pustaka, dan bab metode penelitian(Aisiah & Firza, 2019).

Dengan perkembangannya, universitas bisa dipastikan akan selalu menghasilkan skripsi tiap tahunnya. Namun yang mengkhawatirkan adalah ketidakmampuan dari mahasiswa untuk mengembangkan tema dari hal yang diminati atau dikuasai, (Cahyani & Arif, 2022).

Salah satu cara yang bisa dipakai untuk mengatasi tersebut adalah dengan melakukan temu kembali informasi. *Information retrieval* atau temu kembali informasi merupakan proses menemukan informasi dari teks yang tidak terstruktur untuk memenuhi kebutuhan informasi dari banyak data yang ada di database (Azis et al., 2019; A. Fauzi & Ginabila, 2019). Data yang didapatkan akan diproses menggunakan metode di *text mining*, guna mendapatkan pemahaman lebih dalam tentang isi dokumen, mengidentifikasi pola, atau mengekstraksi pengetahuan yang terkandung dalam teks. *Text mining* saat ini telah menjadi salah satu bidang keilmuan yang sudah tersebar luas guna untuk sebagai cara menganalisis data teks agar menghasilkan informasi yang berguna. (Salloum et al., 2018; Thakur & Kumar, 2022).

Penelitian terdahulu membahas tentang kemiripan dokumen memakai beragam metode, ada yang menggunakan metode clustering (Cahyani & Arif, 2022; Kambey & Dkk, 2020), ada juga yang memakai klasifikasi setelah dilakukan pembobotan dan perhitungan similaritas dengan *cosine similarity* (Efendi & Mustakim, 2017), *Cosine similarity* (Ariantini et al., 2016; Benard Magara et al., 2018; Mawanta et al., 2021; Naf'an et al., 2019; Park et al., 2020; Via & Mumpuni, 2019; Wahid & SN, 2016; Wahyuni et al., 2017), Boyer-Moore (Ahmad et al., 2021), Rabin Karp (Yuliyanti & Rizky, 2020).

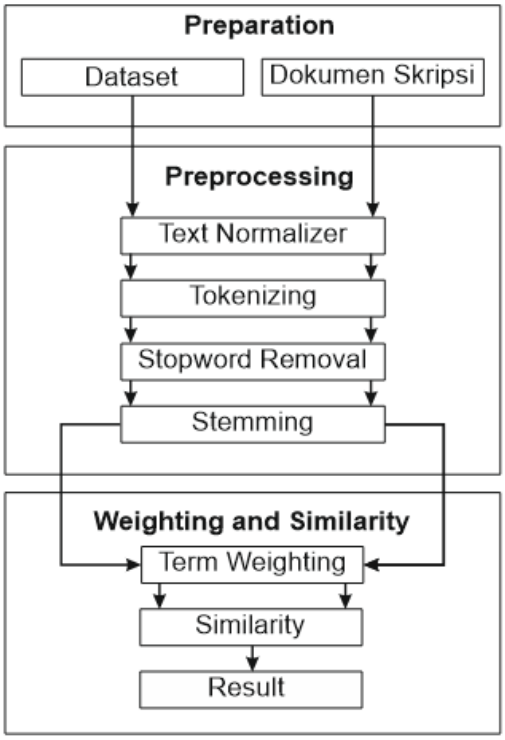
Dari beberapa literatur yang pernah diteliti sebelumnya, *cosine similarity* banyak digunakan dalam menentukan tingkat kemiripan antar teks. Magara (Benard Magara et al., 2018) meneliti algoritma serta kombinasi metrik kesamaan mana yang dapat digunakan guna mengoptimalkan pencarian dan rekomendasi artikel dalam sistem rekomendasi makalah penelitian, *cosine similarity* pun dipilih dibandingkan matrik *similarity* lainnya. Park (Park et al., 2020) dalam penelitiannya menyebut jika akurasi pengklasifikasi konvensional bisa ditingkatkan dengan menambahkan *cosine similarity* sebelum melakukan klasifikasi, lebih lanjut metodologi ini akan disebut sebagai *enhanced classifiers*. Penggunaan pembobotan *term frequency-inverse document frequency* (TF-IDF) juga disarankan jika memakai *cosine similarity*. Pembuatan aplikasi serupa pernah dilakukan oleh Fauzi (R. Fauzi et al., 2022), sistem tersebut dinamai kipcheck, sistem tersebut berhasil untuk memproses hingga 14938 kata untuk sekali input, akurasi yang dihasilkan sistem sebesar 99.88% berdasarkan dengan perhitungan yang dibandingkan dengan perhitungan manual.

Jadi, penelitian ini penulis akan memanfaatkan fungsi perangkingan dari *cosine similarity* untuk membangun sistem analisis kemiripan proposal skripsi. Seperti saran (R. Fauzi et al., 2022), penulis juga akan menambahkan pembobotan kata menggunakan algortima *term frequency-inverse document frequency* sebelum melakukan perhitungan *cosine similarity*.

2. Metode Penelitian

2.1. Arsitektur Perhitungan

Arsitektur dari sistem analisis kemiripan sripsi ini terdiri dari tiga tahap utama yaitu tahap *preparation*, tahap *preprocessing* dan tahap *weighting and similarity*. Arsitektur dari sistem ini dapat dilihat pada gambar 1.



Gambar 1 Arsitektur Perhitungan Sistem

a. Preparation

Persiapan yang perlu dilakukan adalah mempersiapkan dataset. Data yang dipakai adalah judul dan abstrak dari skripsi Fakultas Sains Dan Teknologi Universitas Ibrahimy angkatan 2018. Data mentah berbentuk kumpulan file skripsi yang tersimpan didalam kaset DVD-RW. Data diambil dan disimpan didalam format *Comma Separated Value* (CSV). File CSV dipilih karena mimiliki banyak kelebihan, seperti ringan, ukuran relatif kecil, dan sederhana.

b. *Preprocessing*

Tahap *preprocessing* adalah tahap untuk mengolah dataset yang sudah dipersiapkan ditahap sebelumnya. Tahap ini penting karena data akan dibersihkan supaya hasil yang didapatkan bisa lebih maksimal dan punya kinerja yang baik. Pada tahap *preprocessing* dilakukan *text normalizer* (mengubah text menjadi huruf kecil serta menghilangkan tanda baca dan tanda khusus yang tidak diperlukan), *Tokenizing* (mengubah kalimat yang ada di dalam korpus menjadi unit-unit yang lebih kecil yang disebut token), *stopword removal* (*Stopword* adalah kata-kata umum yang sering muncul dalam teks tetapi tidak memberikan informasi signifikan dalam analisis teks) dan *Stemming* (*Stemming* adalah proses mengubah kata-kata menjadi bentuk dasarnya atau kata dasar (stem) dengan cara menghapus awalan atau akhiran kata).

c. Weighting and Similarity

Tahap ini dilakukan pembobotan terhadap dokumen yang hendak dicari kemiripannya, dokumen yang sebelunya telah melewati proses *preprocessing* akan dihitung bobotnya dan dihitung similaritasnya

2.2. Plagiarisme

Plagiarisme merupakan tindakan mencuri sebuah karya tulis yang mengklaimnya sebagai karya asli sendiri, padahal sebenarnya karya tersebut adalah hasil dari orang lain. Tindakan plagiarisme ini melibatkan penyalinan atau pengambilan bagian-bagian penting dari karya orang lain tanpa memberikan pengakuan atau atribusi yang pantas(Wibowo, 2011).

Hal ini mengarah pada kesalahan dalam menentukan ide, gagasan, dan opini dalam suatu karya. Plagiarisme dianggap sebagai pelanggaran hukum karena merupakan pencurian hak cipta orang lain. Di lingkungan akademik sendiri praktek ini sangat mudah dijumpai, tak terkecuali pada karya tulis ilmiah salah satunya skripsi(Via & Mumpuni, 2019).

2.3. *Information Retrieval*

Information retrieval (IR) merujuk pada proses mencari dan mendapatkan informasi yang relevan dari kumpulan dokumen atau sumber data(Zhou et al., 2012), biasanya disimpan dalam database komputer(Azis et al., 2019). Penggunaan *Information retrieval* system sudah banyak digunakan, penerapan yang paling dirasakan adalah penggunaan mesin pencari atau search engine.

Prinsip dasar penggunaan *information retrieval* system adalah apabila ada sekumpulan dokumen lalu pengguna memasukan pertanyaan, maka sistem akan menampilkan jawaban berupa dokumen yang paling relevan(Azis et al., 2019).

2.4. *Text Mining*

Text mining adalah suatu proses untuk mengekstraksi informasi yang berguna dari teks yang tidak terstruktur. Tujuan utama *text mining* adalah untuk mengubah teks yang didapat dari beragam sumber menjadi bentuk yang dapat dipahami oleh komputer dan kemudian menerapkan metode analisis dan pemodelan untuk mendapatkan informasi

yang berguna dan berharga(Efendi & Mustakim, 2017; Hidayatullah & dkk, 2016; Rosid et al., 2020; Salloum et al., 2018).

2.5. *Term frequency-Inverse document frequency*

TF-IDF (*Term frequency-Inverse document frequency*) adalah sebuah metode yang paling umum digunakan dalam pemrosesan teks untuk menilai kepentingan relatif dari sebuah kata dalam suatu dokumen berdasarkan frekuensi kemunculan (Efendi & Mustakim, 2017; Kang et al., 2020; Wahyuni et al., 2017).

Metode ini menggabungkan dua konsep yaitu *term frequency* dan *inverse document frequency*(Naf'an et al., 2019). Nilai TF-IDF lebih tinggi untuk kata-kata yang sering muncul dalam dokumen tertentu (TF tinggi) tetapi jarang muncul dalam koleksi dokumen secara keseluruhan (IDF tinggi). TF (*term frequency*) menggambarkan seberapa sering sebuah kata muncul dalam sebuah dokumen. Ini mengukur frekuensi kata tersebut dalam dokumen. Semakin tinggi frekuensi kata, semakin penting kata tersebut dalam dokumen tersebut. IDF (*inverse document frequency*) mengukur tingkat keunikan kata dalam seluruh koleksi dokumen. Kata-kata yang jarang muncul di banyak dokumen dianggap memiliki tingkat keunikan yang tinggi dan diberi bobot yang lebih tinggi(Kang et al., 2020; Wahyuni et al., 2017).

Untuk menghitung *term frequency* diperlukan pembagian antara jumlah kemunculan term tersebut dalam sebuah dokumen dengan jumlah keseluruhan term pada dokumen. Frequency term i dalam dokumen j dapat ditulis sebagai berikut:

Persamaan 2-1 Rumus *term frequency*

$$tf_{ij} = \frac{f_{ij}}{max_i(f_{ij})}$$

Kemudian untuk menghitung *inverse document frequency* adalah dengan melakukan perhitungan pembalikan operasi eksponensial basis 10 terhadap hasil dari pembagian seluruh jumlah koleksi dokumen dengan jumlah dokumen yang mengandung term yang dimaksud. Rumus dari *inverse document frequency* didefinisikan sebagai berikut:

Persamaan 2-2 Rumus *inverse document frequency*

$$idf_i = log(\frac{n}{df_i})$$

Rumus untuk menghitung Weight (bobot) W_{ij} adalah dengan rumus TF-IDF seperti yang ditunjukkan pada persamaan 2-6

Persamaan 2-3 Rumus TF-IDF

$$w_{i,j} = tf_{i,j}\ x\ idf_i$$

Jadi, Bobot tertinggi diberikan kepada term yang sering muncul pada dokumen tertentu dan jarang muncul dalam keseluruhan dokumen. Berarti term spesifik itu berada dalam dokume tertentu saja, akan tetapi kemunculannya minim di keseluruhan koleksi dokumen.

2.6. *Cosine Similarity*

Cosine similarity adalah salah satu metode yang sering digunakan dalam analisis teks dan *text mining* untuk mengukur tingkat kemiripan antara dua dokumen(Efendi & Mustakim, 2017; Mawanta et al., 2021; Naf'an et al., 2019; Via & Mumpuni, 2019). *Cosine similarity* tidak terpengaruh pada panjang atau pendeknya suatu dokumen serta tingkat akurasi tinggi. Metode ini didasarkan pada perhitungan sudut kosinus antara dua vektor dalam ruang vektor multidimensi. Hasil dari *cosine similarity* dibatasi antara 0 dan 1, dimana nilai 0 menunjukkan ketidakmiripan dan nilai 1 menunjukkan adanya kemiripan antar dokumen(R. Fauzi et al., 2022; Naf'an et al., 2019). Hal itu karena TF-IDF tidak dapat bernilai negatif dengan sudut antara dua vector tidak dapat lebih dari 90º.

Rumus untuk menghitung *cosine similarity* seperti yang ditunjukkan pada persamaan 2-7.

Persamaan 2-4 Rumus *cosine similarity*

$$\text{Sim}(\text{doc1}, \text{doc2}) = \frac{\sum_{n=1}^{j}(nA\ x\ nB)}{\sqrt{\sum_{n=1}^{j}(nA)^2}\ x\ \sqrt{\sum_{n=1}^{j}(nB)^2}}$$

Nilai j adaalah nilai absolut dari A∩B. nA adalah banyaknya kemunculan kata indeks $(n)^{th}$ dari daftar kata pada kalimat A. Kemudian, nilai nB adalah banyaknya kemunculan kata indeks $(n)^{th}$ dari daftar kata pada kalimat B.

3. Hasil dan Pembahasan

3.1. Persiapan Data

Berikut ini adalah data yang dipersiapkan sebagai dataset, data ini kemudian di unggah kedalam database.

Tabel 1. Dataset

No	Judul	Abstrak
1	sistem informasi keuangan laundry as-syarif menggunakan php & mysql	sistem informasi pun memiliki peranan yang sangat penting dalam suatu instansi, salah satunya juga …
2	pengembangan sistem informasi akademik pada mts salafiyah syafi'iyah menggunakan framework codeigniter dan mysql	salah satu lembaga yang berada dalam naungan pondok pesantren salafiyah syafi'iyah adalah madrasah …
3	pengembangan sistem informasi akademik pada mts salafiyah syafi'iyah menggunakan framework codeigniter dan mysql	salah satu lembaga yang berada dalam naungan pondok pesantren salafiyah syafi'iyah adalah madrasah …
…	…	…
14	pengembangan sistem informasi barokah pegawai pada instansi fakultas tarbiyah berbasis web	fakultas tarbiyah universitas ibrahimy sebagai salah satu tertua fakultas di universitas …

3.2. Skenario Pengujian

a. Skenario Pertama

Skenario ini menggunakan masukan data yang mirip sepenuhnya dengan data yang ada di data set.

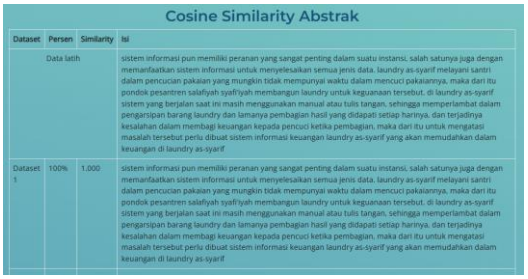
Query dari judul dan proposal skripsi yang di masukkan “sistem informasi keuangan laundry as-syarif menggunakan php & mysql” dan ”sistem informasi pun memiliki peranan yang sangat penting dalam suatu instansi, salah satunya juga …”

Hasilnya sebagai berikut:

Cosine Similarity Judul

Dataset	Persen	Similarity	Isi
Data latih			sistem informasi keuangan laundry as-syarif menggunakan php & mysql
Dataset 1	100%	1.000	sistem informasi keuangan laundry as-syarif menggunakan php & mysql
Dataset 8	26.7%	0.267	sistem informasi pengelolaan keuangan asrama bahasa menggunakan php & mysql
Dataset 11	11%	0.110	sistem informasi rekapitulasi skor pelanggaran siswa madrasah i'dadiyah menggunakan php dan mysql

Gambar 2 Hasil percobaan pertama judul



Gambar 3 Hasil percobaan pertama abstrak

b. Skenario Kedua

Skenario kedua ini menggunakan masukan data yang mirip sebagian dengan data yang ada di data set.

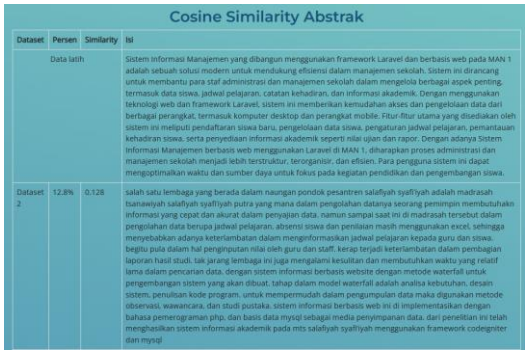
Query dari judul dan proposal skripsi yang di masukkan ”rancang bangun sistem informasi manajemen pada MAN 1 menggunakan laravel berbasis website” dan ”Sistem Informasi Manajemen yang dibangun menggunakan framework Laravel dan berbasis web pada MAN 1 adalah sebuah solusi modern untuk mendukung efisiensi dalam manajemen sekolah. Sistem ini dirancang untuk membantu para staf administrasi dan…”

Hasilnya sebagai berikut:

Cosine Similarity Judul

Dataset	Persen	Similarity	Isi
Data latih			rancang bangun sistem informasi manajemen pada MAN 1 menggunakan laravel berbasis website
Dataset 6	18.4%	0.184	sistem informasi penilaian siswa peserta pkl pada smk ibrahimy 1 sukorejo berbasis website
Dataset 3	11.4%	0.114	sistem informasi manajemen al-risalah berbasis web menggunakan php dan mysql pada madrasah aliyah salafiyah syafi'iyah putri (masspi)
Dataset 1	2.1%	0.021	sistem informasi keuangan laundry as-syarif menggunakan php & mysql

Gambar 4 Hasil percobaan kedua judul



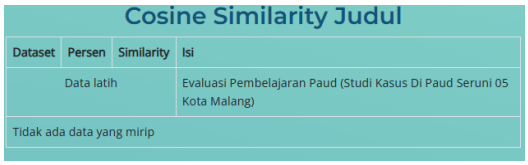
Gambar 5 Hasil percobaan kedua abstrak

c. Skenario Ketiga

Skenario ketiga ini menggunakan masukan data yang sama sekali berbeda dengan data yang ada di data set

Query dari judul dan proposal skripsi yang di masukkan " Evaluasi Pembelajaran Paud (Studi Kasus Di Paud Seruni 05 Kota Malang)" dan "Penelitian ini merupakan studi kasus yang bertujuan untuk mengevaluasi proses pembelajaran di Pendidikan Anak Usia Dini (PAUD) Seruni 05 Kota Malang. Evaluasi pembelajaran di PAUD memiliki peran penting dalam meningkatkan kualitas pendidikan …"

Hasilnya sebagai berikut:



Gambar 6 Hasil percobaan ketiga judul

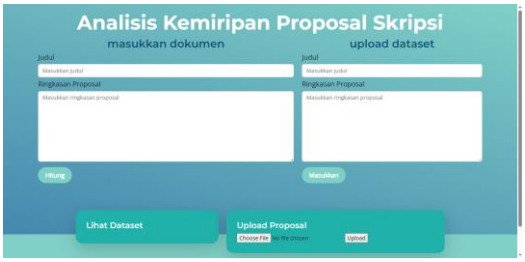
Gambar 7 Hasil percobaan ketiga abstrak

3.3. Analisis Hasil

Dari hasil percobaan yang telah dilakukan dapat dilihat bahwa sistem bisa menemukan kemiripan antara dua dokumen. Pada percobaan pertama didapatkan hasil kemiripan sebesar 100%, dimana masukan yang di berikan benar benar mirip dengan data yang ada di dataset. Pada percobaan kedua, didapatkan kemiripan sebesar 18.4%, dimana data masukan disesuaikan agar tidak terlalu mirip dengan dataset, hasil menunjukkan bahwa persentase termasuk kedalam tahap wajar. Pada percobaan terakhir, sistem diuji dengan data yang berbeda dengan yang ada di dataset, hasilnya sistem memberikan rangking kemiripan terhadap sedikit sekali dataset, dan bahkan ada yang tidak memiliki kemiripan sama sekali. Hal ini menunjukkan bahwa apabila masukan benar benar berbeda sistem akan dengan sendirinya mengenali bahwa masukan tersebut tidak ada kemiripannya dengan dataset.

3.4. Implementasi Sistem

Berikut ini adalah implementasi dari sistem analisis kemiripan prosposal skripsi



Gambar 8 Tampilan sistem

Sistem ini menerima masukan berupa judul dan ringkasan dari proposal skripsi. Untuk mengunggah dataset bisa dilakukan satu persatu, atau bisa menggunakan format *Comma-Separated Values* (CSV) untuk unggah dataset dalam jumlah besar.

4. Kesimpulan

Berdasarkan pembahasan yang telah dilakukan dalam pengembangan sistem analisis, simpulan yang dihasilkan adalah sistem berhasil menghitung persentase kemiripan dari proposal skripsi yang dimasukkan kedalam sistem dan sistem yang dibangun, sistem mendeteksi kemiripan sebesar 100% apabila masukan dari pengguna sama persis dengan yang ada di dataset. Selain itu, sistem bisa memberikan gambaran bagi panitia skripsi tentang penelitian yang akan dilakukan.

Beberapa saran bagi penelitian lebih lanjut adalah menambahkan fitur atau metode untuk mengelola sinonim dan akronim, sehingga dapat meningkatkan

kemampuan sistem serta bisa menambahkan evaluasi terhadap sistem, seperti menggunakan Normalized Discounted Cumulative Gain, Mean Reciprocal Rank, atau Mean Average Precision. Sehingga sistem dapat dievaluasi sejauh mana persentase keakuratan yang dihasilkan.

Referensi

Ahmad, I., Borman, R. I., Caksana, G. G., & Fakhrurozi, J. (2021). Implementasi String Matching Dengan Algoritma Boyer-Moore Untuk Menentukan Tingkat Kemiripan Pada Pengajuan Judul Skripsi/TA Mahasiswa (Studi Kasus: Universitas XYZ). *SINTECH (Science and Information Technology) Journal*, *4*(1), 53–58. https://doi.org/10.31598/sintechjournal.v4i1.699

Aisiah, A., & Firza, F. (2019). Kendala yang Dihadapi Mahasiswa Jurusan Sejarah dalam Menulis Proposal Skripsi. *Diakronika*, *18*(2), 90. https://doi.org/10.24036/diakronika/vol18-iss2/70

Ariantini, D. A. R., Lumenta, A. S. M., & Jacobus, A. (2016). Pengukuran Kemiripan Dokumen Teks Bahasa Indonesia Menggunakan Metode Cosine Similarity. *Jurnal Teknik Informatika*, *9*(1), 1–8. https://doi.org/10.35793/jti.9.1.2016.13752

Azis, M. A., Hamid, A., Fauzi, A., Yudhistira, Yulianto, E., Riyanto, V., Ridwansyah, & Sfenrianto. (2019). Information retrieval system in text-based skripsi document search file using vector space model method. *Journal of Physics: Conference Series*, *1367*(1), 012016. https://doi.org/10.1088/1742-6596/1367/1/012016

Benard Magara, M., Ojo, S. O., & Zuva, T. (2018). A comparative analysis of text similarity measures and algorithms in research paper recommender systems. *2018 Conference on Information Communications Technology and Society (ICTAS)*, 1–5. https://doi.org/10.1109/ICTAS.2018.8368766

Cahyani, L., & Arif, M. (2022). Text Mining untuk Pengelompokan Skripsi di Prodi Pendidikan Informatika Universitas Trunojoyo Madura. *Jurnal Ilmiah Edutic : Pendidikan Dan Informatika*, *8*(2), 97–108. https://doi.org/10.21107/edutic.v8i2.13020

Efendi, Z., & Mustakim, M. (2017). Text Mining Classification sebagai Rekomendasi Dosen Pembimbing Tugas Akhir Program Studi Sistem Informasi. *Seminar Nasional Teknologi Informasi Komunikasi Dan Industri*, *0*(0), 235–242. http://ejournal.uin-suska.ac.id/index.php/SNTIKI/article/view/3273

Fauzi, A., & Ginabila, G. (2019). Information Retrieval System Pada File Pencarian Dokumen Tesis Berbasis Text Menggunakan Metode Vector Space Model. *Jurnal Pilar Nusa Mandiri*, *15*(1), 41–46. https://doi.org/10.33480/pilar.v15i1.61

Fauzi, R., Iqbal, M., & Haryanti, T. (2022). Design and Implementation of a Final Project Plagiarism Detection System Using Cosine Similarity Method. *IJAIT (International Journal of Applied Information Technology)*, *05*(02), 1. https://doi.org/10.25124/ijait.v5i02.4146

Hadi, S. (2016). Pemeriksaan Keabsahan Data Penelitian Kualitatif Pada Skripsi [Examination of the Validity of Qualitative Research Data on Thesis]. *Ilmu Pendidikan*, *22*(1), 21–22. https://doi.org/http://dx.doi.org/10.17977/jip.v22i1.8721

Hidayatullah, F., & dkk. (2016). Penerapan Text Mining dalam Klasifikasi Judul Skripsi. *Seminar Nasional Aplikasi Teknologi Informasi (SNATi) Agustus*, 1907–5022. https://journal.uii.ac.id/Snati/article/view/6232

Kambey, G. E. I., & Dkk. (2020). Penerapan Clustering pada Aplikasi Pendeteksi Kemiripan Dokumen Teks Bahasa Indonesia. *Penerapan Clustering Pada Aplikasi Pendeteksi Kemiripan Dokumen Teks Bahasa Indonesia*, *15*(2), 75–82. https://ejournal.unsrat.ac.id/v3/index.php/informatika/article/view/28907

Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, *7*(2), 139–

172. https://doi.org/10.1080/23270012.2020.1756939

Mawanta, I., Gunawan, T. S., & Wanayumini, W. (2021). Uji Kemiripan Kalimat Judul Tugas Akhir dengan Metode Cosine Similarity dan Pembobotan TF-IDF. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(2), 726. https://doi.org/10.30865/mib.v5i2.2935

Naf'an, M. Z., Burhanuddin, A., & Riyani, A. (2019). Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen. *Jurnal Linguistik Komputasional (JLK)*, *2*(1), 23. https://doi.org/10.26418/jlk.v2i1.17

Park, K., Hong, J. S., & Kim, W. (2020). A Methodology Combining Cosine Similarity with Classifier for Text Classification. *Applied Artificial Intelligence*, *34*(5), 396–411. https://doi.org/10.1080/08839514.2020.1723868

Rosid, M. A., Fitrani, A. S., Astutik, I. R. I., Mulloh, N. I., & Gozali, H. A. (2020). Improving Text Preprocessing For Student Complaint Document Classification Using Sastrawi. *IOP Conference Series: Materials Science and Engineering*, *874*(1), 012017. https://doi.org/10.1088/1757-899X/874/1/012017

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2018). Using Text Mining Techniques for Extracting Information from Research Articles. In *Studies in Computational Intelligence* (Vol. 740, pp. 373–397). https://doi.org/10.1007/978-3-319-67056-0_18

Thakur, K., & Kumar, V. (2022). Application of Text Mining Techniques on Scholarly Research Articles: Methods and Tools. *New Review of Academic Librarianship*, *28*(3), 279–302. https://doi.org/10.1080/13614533.2021.1918190

Via, Y. vita, & Mumpuni, R. (2019). Deteksi Kemiripan Dokumen Publikasi Skripsi Mahasiswa Menggunakan Algoritma Modifikasi Cosine Similarity. *Journal of Information Engineering and Educational Technology*, *3*(2), 57–61. https://doi.org/10.26740/jieet.v3n2.p57-61

Wahid, D. H., & SN, A. (2016). Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *10*(2), 207. https://doi.org/10.22146/ijccs.16625

Wahyuni, R. T., Prastiyo, D., & Supraptono, E. (2017). Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi. *Teknik Elektro*, *9*(1). https://doi.org/10.15294/jte.v9i1.10955

Wibowo, A. (2011). *Mencegah dan Menanggulangi Plagiarisme di Dunia Pendidikan*. 195–200. https://journal.fkm.ui.ac.id/kesmas/article/view/84/85

Yuliyanti, S., & Rizky. (2020). Implementasi Algoritma Rabin Karp Untuk Mendeteksi Kemiripan Dokumen Stmik Bandung. *Jurnal Bangkit Indonesia*, *10*(02), 1. https://doi.org/10.52771/bangkitindonesia.v10i02.124

Zhou, H., Liu, B., & Liu, J. (2012). Research on Mechanism of the Information Retrieval Based on Ontology Label. *Procedia Engineering*, *29*, 4259–4266. https://doi.org/10.1016/j.proeng.2012.01.654