Sistem Penentuan Kemiripan Antar Skripsi Menggunakan Metode *Cosine Similarity* Pada Perpustakaan

Destri Natalia Lindang[1)], Andi Yulia Muniar[2)], Agus Halid[3)], Muhajirin[4)], Amran Amiruddin[5)]
[1,2,3,4,5] Universitas Teknologi AKBA Makassar
destri16@mhs.akba.ac.id[1)], andiyulia@akba.ac.id[2)], agushalid@akba.ac.id[3)], aji@akba.ac.id[4)], amran@akba.ac.id[5)]

Abstak

Pada penentuan skripsi memiliki beberapa kendala sehingga menyebabkan sulit nya memilih skripsi yang berkualitas sehingga penelitian ini bertujuan merancang dan menerapkan sebuah sistem penentuan kemiripan antar skripsi dengan metode cosine similarity pada perpustakaan untuk mempermudah mahasiswa dalam mencari kemiripan antar skripsi. Data diperoleh melalui penelitian Pustaka. Perancangan sistem penentuan kemiripan antar skripsi ini menggunakan bahasa pemrograman PHP dan *database* yang digunakan yaitu MySQL. Hasil dari penelitian ini menunjukkan bahwa sistem penentuan kemiripan antar skripsi dapat diimplementasikan dan dinyatakan layak digunakan karena hasil persentase pengujian akurasi memiliki rata-rata sebesar 97% dengan waktu komputasi yaitu 0,277154 detik.

Keywords: *Sistem informasi, Skripsi, Cosine Similarity, Perpustakaan.*

I. PENDAHULUAN

Akses informasi dapat diperoleh dengan mudah seiring dengan perkembangan teknologi dalam bidang komputasi dan telekomunikasi. Informasi tersebut dapat berbentuk teks, gambar, suara, ataupun video. Sistem informasi berbasis teks merupakan informasi yang disimpan dalam dokumen berupa huruf ataupun angka. Dokumen yang ada di sistem informasi dapat diketahui bilamana telah terbuka. Bilamana butuh informasi, maka file yang dibutuhkan dibuka satu persatu. Hal ini akan menyulitkan pengguna dalam menghasilkan informasi yang cepat dan tepat. Olehnya itu, diperlukan suatu cara agar pengguna dapat mengakses informasi secara cepat dan tepat.

Tulisan ilmiah dapat ditemukan dengan mudah dengan melihat tingkat relevansi atau kecocokan suatu dokumen dengan dokumen lainnya. Manusia dapat dengan mudah menentukan tingkat relevansi suatu dokumen dengan menggunakan *keyword*. Dengan mengetikkan kata kunci, sistem informasi akan menampilkan dokumen-dokumen tersebut.

Pencarian judul skripsi untuk menyelesaikan tugas akhir dari mahasiswa merupakan sesuatu yang dapat dikatakan sulit dan juga dapat dikatakan mudah. Banyaknya judul yang sudah ada dan banyaknya kemiripan judul dapat diketahui melalui seleksi dari para dosen dan tim evaluasi proposal. Dengan adanya judul yang sama, tidak menutup kemungkinan isi dari judul skripsi tersebut juga sama, sehingga dapat menimbulkan tindak plagiarisme.

Perpustakaan sebagai tempat penyimpanan skripsi untuk menunjang kegiatan belajar mahasiswa setiap hari. Jika pada suatu perpustakaan tidak adanya sistem yang menentukan kemiripan antar skripsi pada perpustakaan, mahasiswa melakukannya dengan cara membuka atau membaca satu persatu skripsi yang ada untuk dilihat kemiripannya. Cara seperti itu akan membutuhkan waktu yang lama untuk dapat melihat kemiripan antar skripsi, sehingga mahasiswa akan merasa kesulitan dalam mencari dan mengetahui kemiripan antar skripsi.

II. LANDASAN TEORI

Penelitian yang dilakukan oleh (Apriyanto & Aribowo, 2018) telah diusulkan sebuah metode untuk merancang suatu aplikasi atau sistem yang dapat memberikan kemudahan dosen ataupun mahasiswa dalam melakukan pengecekan judul skripsi yang diajukan dengan skripsi-skripsi terdahulu. Dari penelitian yang dilakukan menghasilkan aplikasi pengecekan kemiripan judul skripsi di Teknik Informatika Universitas Ahmad Dahlan (UAD) dengan metode *Usability Test*, 0% responden menyatakan penilaian sistem tidak diterima, 20% responden memberikan penilaian marginal dan 80% responden menyatakan sistem bisa diterima.

Berdasarkan penelitian yang dilakukan (Melita, Amrizal, Suseno, & Dirjam, 2018) diusulkan metode *similarity* yang dapat digunakan untuk melakukan pencarian dokumen relevan dengan yang diinginkan. Metode *similarity* yang digunakan yaitu *cosine similarity* menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF) dan menerapkan teks preprocessing terlebih dahulu. Teks *preprocessing* meliputi *tokenizing*, *stopword* removal atau *filtering* dan *stemming*. Hasil uji coba dengan pengujian *confusion* matrix didapatkan: recall 88,7%, *precision* 100%, *accuracy* 88,73%, dan *error rate* 11,27%.

Berdasarkan penelitian (Herwijayanti, Ratnawati, & Mufllikhah, 2018), penerapan klasifikasi berita *online* dengan menggunakan tf-idf dan *cosine similarity*, memerlukan proses *preprocessing* yaitu *tokenizing*, *stopward* dan *stemming* dapat mempercepat proses *cosine similarity*. Tujuannya adalah untuk mengurangi *human error* serta mengurangi terjadinya kesalahan pengkategorian. Klasifikasi mampu mengelompokkan berita dengan tingkat akurasi sebesar 91,25%.

Penelitian (Apriyanto & Aribowo, 2018), memaparkan bahwa fungsi *similarity* adalah fungsi yang

menerima dua buah objek dan mengembalikan nilai kemiripan (*similarity*) antara kedua objek tersebut berupa bilangan riil. Semakin besar hasil fungsi *similarity*, maka kedua objek akan dianggap semakin mirip. Sebaliknya, semakin kecil hasil *similarity*, maka kedua objek tersebut dianggap semakin berbeda. *Cosine similarity* adalah perhitungan kesamaan antara dua vektor n dimensi dengan mencari kosinus dari sudut diantara keduanya dan sering digunakan untuk membandingkan dokumen dalam *text mining*. Rumus *cosine similarity* adalah sebagai berikut:

$$Similarity\ (x, y) = cos(\theta) = \frac{x, y}{||x||||y||}$$

Dimana :

x.y : *vector dot product* dari x dan y, dihitung dengan $\sum_{k=1}^{n} \boldsymbol{x_k\, y_k}$;

‖x‖ : panjang vektor x, dihitung dengan $\sum_{k=1}^{n} \boldsymbol{x_k^2}$;

‖y‖ : panjang vektor y, dihitung dengan $\sum_{k=1}^{n} \boldsymbol{y_k^2}$

Metode TF-IDF menggabungkan dua konsep untuk perhitungan bobot, yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen tertentu dan inverse frekuensi dokumen yang mengandung kata tersebut. Frekuensi kemunculan kata pada suatu dokumen yang diberikan menunjukkan seberapa penting kata itu di dalam dokumen tersebut. Frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. Sehingga bobot hubungan antara sebuah kata dan sebuah dokumen akan tinggi apabila frekuensi kata tersebut tinggi di dalam dokumen dan frekuensi keseluruhan dokumen yang mengandung kata tersebut yang rendah pada kumpulan dokumen. (Wahyuni et al., 2017).

1. *Term Frequency* (TF)

Term Frequency (TF) adalah frekuensi dari kemunculan sebuah term dalam dokumen yang bersangkutan. Menyatakan jumlah keberadaan suatu *term* atau kata dalam sebuah dokumen, perhitungan bobot term sebagai berikut:

$$q = (tf) * (idf)$$

Keterangan:

q : Nilai bobot term;

tf : Nilai term frequency;

idf : Nilai inverse document frequency

2. *Inverse Document Frequency* (IDF)

Inverse Document Frequency (IDF) merupakan sebuah perhitungan dari bagaimana term didistribusikan secara luas pada koleksi dokumen yang bersangkutan. Berfungsi mengurangi bobot suatu *term* jika kemunculannya tersebar di seluruh koleksi dokumen dirumuskan dengan persamaan:

$$IDF = \log\left(\frac{\boldsymbol{n}}{\boldsymbol{df}}\right)$$

Keterangan:

IDF : Nilai inverse document frequency;

n : Jumlah dokumen di dalam koleksi;

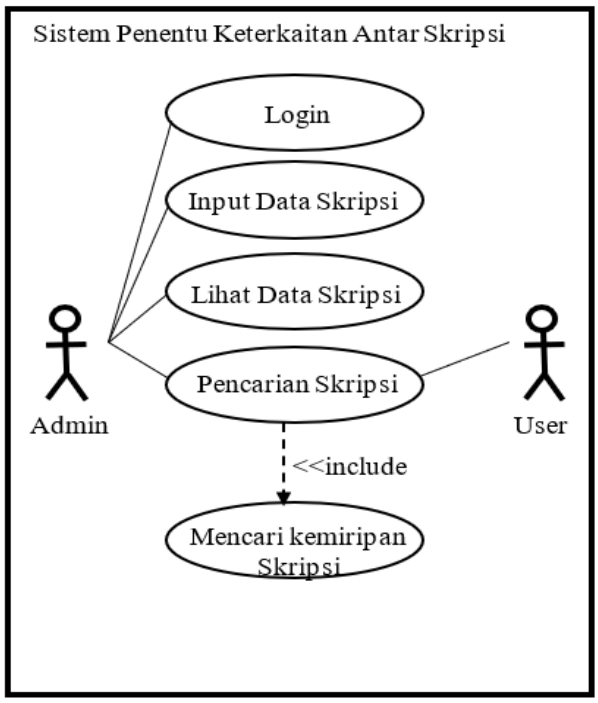
df : Nilai document frequency

III. METODE PENELITIAN

Sistem dirancang berdasarkan sistem penentu kemiripan antar skripsi berdasarkan kata kunci. Pada saat pencarian referensi skripsi terdahulu, maka yang harus dilakukan adalah mencari sendiri kebutuhan di perpustakaan. Proses pencarian ini akan membutuhkan waktu sehingga dibutuhkan sistem penentuan kemiripan antar skripsi berdasarkan kata kunci. Berdasarkan nilai akurasinya, akan diketahui seberapa besar tingkat kemiripan antar skripsi dengan kata kunci yang diberikan.

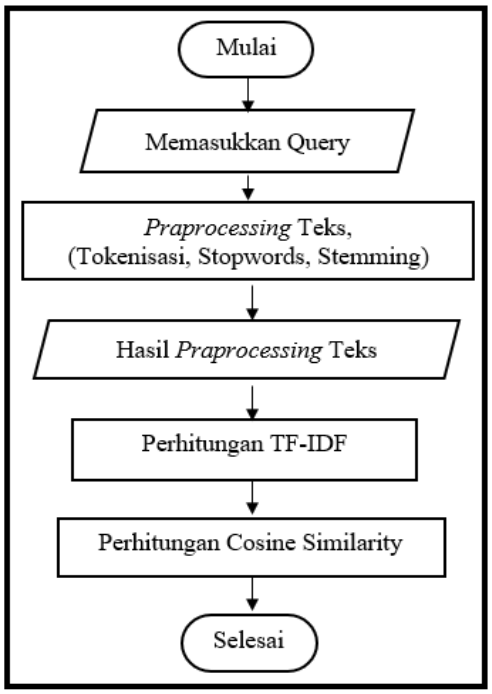
Proses pengujian dilakukan dengan mengambil data abstrak, kesimpulan dan saran pada dokumen skripsi. Kemudian data diolah dengan menggunakan metode *cosine similarity*, dengan menggunakan pembobotan TF-IDF.

Perolehan data diambil melalui penelitian pustaka dengan mempelajari, memahami dan mengutip teori-teori dari buku, jurnal atau karya tulis lainnya yang membahas mengenai topik yang diangkat oleh peneliti. Perancangan sistem memberikan gambaran sistem yang akan dibangun atau dikembangkan, serta untuk memahami alur informasi dan proses sebagaimana diperlihatkan Gambar 1.



Gambar 1. *Use Case Diagram*

Gambar 1 memperlihatkan bagaimana pengguna melakukan pencarian kemiripan dari skripsi yang akan dibuatnya dengan menggunakan metode *cosine similarity*. Penelitian terdahulu akan diinput oleh admin untuk memenuhi kebutuhan pencocokan.



Gambar 2. *Flowchart* Proses

Flowchart sistem sebagaimana Gambar 2, dimulai dengan *query* terhadap skripsi yang dijadikan sebagai referensi. Setelah itu, dilakukan praprosessing terhadap teks. Hasil yang diperoleh dari pra-prosessing teks, diolah dengan TF-IDF yang selanjutnya diproses menggunakan metode *cosine similarity*.

IV. HASIL DAN PEMBAHASAN

Pengujian sistem diperlukan untuk memenuhi kebutuhan pengguna agar dapat dipastikan pengguna bebas dari kesalahan, dalam hal ini digunakan metode *blackbox testing*. Pengujian ini mengutamakan pada fungsionalitas yang ada di setiap bagian sistem tanpa mengetahui *source code*. Tujuan dari pengujian ini yaitu untuk memastikan setiap bagian sudah sesuai dengan alur proses yang ditetapkan dan memastikan semua kesalahan masukan yang dilakukan oleh pengguna dapat ditangani oleh sistem. Pemilihan metode ini didasarkan bahwa pengujian tidak membutuhkan waktu yang cukup lama, dan bisa mempersiapkan spesifikasi pengujian saat dilakukan analisa sistem secara bersamaan

Berdasarkan pengujian dengan *Confusion Matrix*, yakni proses perbandingan antara dataset dari hasil klasifikasi data sebenarnya dengan jumlah data secara keseluruhan. *Confusion Matrix* berbentuk tabel matriks yang menggambarkan kinerja model klasifikasi pada serangkaian data uji yang nilai sebenarnya diketahui.

Tabel 1. *Confusion Matrix*

	Relevan	**Tidak Relevan**
Di-Retrieve	TP (True Positive)	FP (False Positive)
Tidak Di-Retrieve	FN (False Negative)	TN (True Negative)

Pada Tabel 1 terdapat 4 istilah sebagai representasi hasil proses klasifikasi pada *confusion matrix*. Keempat istilah tersebut adalah *True Positive* (TP) merupakan data positif yang diprediksi benar. *False Positive* (FP) merupakan data negatif namun diprediksi sebagai data positif. *False Negative* (FN) merupakan data positif namun diprediksi sebagai data negative. Dan *True Negative* (TN) merupakan data negatif yang diprediksi benar.

Tabel 2 *Ground Truth Table*

ID	**Keyword**	**ID Skripsi yang Relevan**
q1	Metode Certainty Factor	24,25,45
q2	Metode Cosine Similarity	27,54,82,83,84
q3	Rancang Bangun Aplikasi	23,37,38,39,51,68,91, 92
q4	Sistem Pakar	24,25,26,45,98
q5	Metode TF IDF	15,27,28,55
q6	Data Mining	64,66,67
q7	Augmented Reality	16,17,18,19,20,50
q8	Sistem Pendukung Keputusan	32,72,107
q9	Jaringan Saraf Tiruan	14,29,30,31,58
q10	Mikrokontroler Arduino Uno	38,40,41,42,44,59

Proses selanjutnya adalah melakukan evaluasi pencarian berdasarkan masing-masing keyword. Untuk menentukan nilai masing-masing TP, FP, FN dan TN. Kemudian hitung nilai *precision*, recall, F1-measure, dan *accuracy*.

Tabel 3 Matriks Pengukuran Relevansi

q	Confusion Matrix				Relevansi			
	TP	FP	FN	TN	Pre	Re	F-me	Ac
1	3	0	0	97	1,00	1,00	1,00	1,00
2	5	3	0	92	0,63	1,00	0,77	0,97
3	5	3	3	89	0,63	0,63	0,63	0,94
4	5	0	0	95	1,00	1,00	1,00	1,00
5	3	0	1	96	1,00	0,75	0,86	0,99
6	3	5	0	92	0,38	1,00	0,55	0,95
7	6	0	0	94	1,00	1,00	1,00	1,00
8	3	5	0	92	0,38	1,00	0,55	0,95
9	5	2	0	93	0,71	1,00	0,83	0,98
10	5	3	1	91	0,63	0,83	0,71	0,96
Rata-Rata					0,74	0,92	0,79	0,97

Tabel 3 dilakukan evaluasi hasil pencarian dengan membatasi jumlah hasil pencarian N=8. Setelah mendapatkan matriks pengukuran relevansi langkah

selanjutnya adalah menghitung nilai rata-rata akhir dari masing-masing pencarian untuk mengetahui tingkat akurasinya.

Tabel 4 Evaluasi Hasil Pengujian

N=	Precision	Recall	Fmeasure	Accuracy (%)
1	1	0,23	0,37	96%
2	0,95	0,43	0,58	97%
3	0,97	0,66	0,77	98%
4	0,9	0,75	0,8	98%
5	0,86	0,85	0,84	98%
6	0,82	0,9	0,84	98%
7	0,77	0,92	0,82	98%
8	0,74	0,92	0,79	97%
9	0,73	0,93	0,79	97%
Rata-Rata akurasi				97%

Berdasarkan hasil pengujian akurasi Tabel 4 dengan data pengujian sebanyak 10 query dan dilakukan sebanyak 9 kali pengujian, diperoleh tingkat akurasi sebesar 97%. Dengan rata-rata waktu komputasi 0,277154 detik.

V. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan dapat disimpulkan sebagai berikut :

1. Sistem penentuan kemiripan antar skripsi ini menggunakan metode *cosine similarity* untuk menentukan tingkat kemiripan antar skripsi.
2. Berdasarkan hasil penelitian serta pengujian sistem menggunakan pengujian *cosine similarity* dengan *keyword* pengujian sebanyak 10 kali dan evaluasi n sebanyak 9 kali, maka didapatkan nilai rata-rata akurasi yaitu sebesar 97% dan rata-rata waktu komputasi yaitu 0,277154 detik.

UCAPAN TERIMA KASIH

Ucapan terima kasih ditujukan kepada pimpinan dan seluruh staf Universitas Teknologi AKBA Makassar yang telah memberikan kesempatan melaksanakan kegiatan penelitian khususnya pada bagian Perpustakaan.

REFERENSI

[1] Apriyanto, Ibnu, Aribowo. 2018. Rancang Bangun Aplikasi Pengecekan Kemiripan Judul Skripsi Dengan Metode Cosine Similarity Studi Kasus Program Studi Teknik Informatika UAD. Sarjana Teknik Informatika. 6(2): 43—52.

[2] Herwijayanti, B., Ratnawati, D. E. and Muflikhah, L. (2018). Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. 2(1), 306–312

[3] Melita, Ria, dkk. 2018. Penerapan Metode Term Frequency Inverse Document Frequency (TF-IDF) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web Studi Kasus Syarah Umdatil Ahkam. Teknik Informatika. 11(2): 149-164.

[4] Wahyuni, Rizki, Dhidik, Eko. 2017. Penerapan Algoritma Cosine Similarity Dan Pembobotan TF-IDF Pada Sistem Klasifikasi Dokumen Skripsi. Teknik Elektro. 9(1): 18-23.