e-ISBN 978-623-93343-4-5
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 15 September 2021

Implementasi *Text Mining* Klasifikasi Topik Tugas Akhir Mahasiswa Teknik Informatika Menggunakan Pembobotan TF-IDF dan Metode *Cosine Similarity* Berbasis Web

Nila Andriani[1], Arief Wibowo[2]
Fakultas Teknologi Informasi
Universitas Budi Luhur Jakarta
nilaandriani89@gmail.com
Jl Ciledug Raya, Petukangan Utara, Jakarta Selatan, DKI Jakarta, 12260, Indonesia

Abstrak. Pembuatan tugas akhir sebelumnya perlu ditentukan topik yang diambil. Masalah dalam penentuan topik sebuah skripsi adalah diperlukan pemahaman terkait isi dokumen dikarenakan topik dalam skripsi terdapat kategori berbeda, maka diperlukan sistem yang dapat melakukan pengklasifikasian agar pengetahuan terkait topik diketahui dengan cepat dan tepat. Implementasi dari penelitian ini bertujuan untuk dapat mengkategorikan topik skripsi secara komputerisasi atau otomatis menggunakan TF-IDF dan metode Cosine Similarity serta penggunaan sample data skripsi mahasiswa Teknik Informatika Universitas Budi Luhur. Selanjutnya data tersebut akan diproses dengan perhitungan mencari bobot terhadap kata dengan mengalikan banyaknya kata yang dicari dengan Inversed Document Frequency. Setelah perhitungan TF/IDF dan menemukan nilai bobotnya, kemudian menghitung similaritas dengan metode cosine similarity. Hasil dari perhitungan cosine similarity ditampilkan dalam bentuk persentase. Berdasarkan hasil pengujian pada datalatih dan data uji menghasilkan persentase sebesar 86,66%, dengan demikian disimpulkan bahwa metode cosine similarity mendeteksi tingkat similaritas dengan hasil yang cukup baik dan tepat.

Kata Kunci: Algoritma *Cosine Similarity*, Perhitungan TF-IDF, Dokumen Skripsi

1 Pendahuluan

Dalam kemajuan teknologi yang begitu pesat terutama untuk menyajikan sebuah informasi. Manusia membutuhkan bantuan komputer, karena komputer mempunyai banyak kelebihan yakni dalam hal kecepatan kemudian mengenai akurasi serta efisen dalam mengolah data jika dibandingkan oleh manual system. Kelebihan komputer dalam melakukan pengolahan data menjadi sebuah atau kumpulan informasi, cepatnya perkembangan komputer di berbagai aspek, seperti pada dunia bisnis serta dunia pendidikan. Sistem pengolahan data yang baik akan dapat berfungsi mengatasi permasalahan yang terjadi serta tentunya dapat menghasilkan informasi secara dengan cepat, tepat dan memiliki nilai akurasi tinggi. Fungsi lainya dapat digunakan untuk menyusun studi literatur dari sebuah penelitian. Studi literatur adalah bagian dari proses penelitian dengan melakukan pengumpulan referensi bacaan terkait permasalahan serta tujuan penelitian [1].

Untuk mengklasifikasikan dokumen berdasarkan topik yang telah ditentukan, maka dibutuhkan aplikasi yang bisa mendeteksi nilai similarity dari dokumen tugas akhir mahasiswa Budi Luhur. Sistem ini difokuskan untuk pengklasifikasian, sehingga dokumen yang diperiksa dari sistem ini adalah dokumen judul dari skripsi mahasiswa Teknik Informatika Universitas Budi Luhur yang sudah dipublikasi ke perpustakaan dijadikan sebagai data latih dari penelitian ini.

Pada aplikasi yang memiliki jumlah dokumen banyak serta proses bertambahnya data dengan cepat memang diperlukannya aplikasi yang dapat melakukan proses pengklasifikasian teks secara otomatis. Dalam proses penggolongan teks terdapat dua cara, yakni klasterisasi teks dan megklasifikasikan teks. Klasterisasi teks merupakan cara yang saling terhubung dengan cara menemukan data yang memang belum terkelompokkan (*unsupervised*) dari banyaknya jumlah dokumen.

Dan untuk cara mengklasifikasikan teks merupakan proses untuk membentuk golongan atau kelas dari sebuah dokumen yang pada dasarnya kategori atau kelompok sudah ditetapkan atau diketahui (*supervised*) [2]. Proses mengklasifikasikan dokumen, terutama pada *electronic document* yang jumlahnya cukup banyak memang dibutuhkan agar kumpulan data tersebut dapat diolah atau diproses menjadi sebuah atau kumpulan informasi yang benar. Proses mengklasifikasikan ini merupakan upaya dalam mengategorikan dokumen berdasar pada kategori serta ciri yang telah ditetapkan. Bila dokumen yang akan diproses berjumlah sangat banyak dilakukan secara manual akan memerlukan atau memakan banyak wakktu dan tenaga sumber daya manusai yang memproses. Maka

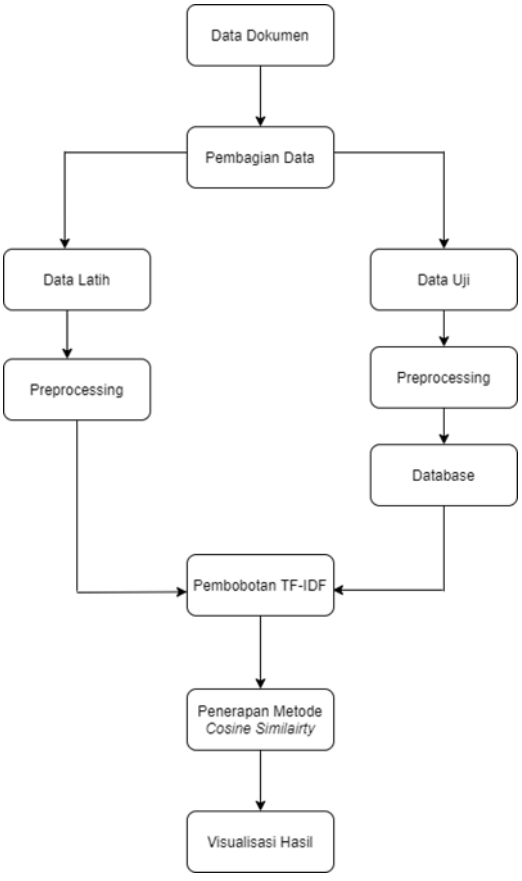
metode yang bisa dipergunakan untuk melakukan proses mengklasifikasikan dokumen yang banyak tersebut secara otomatis atau terkomputerisasi dengan aplikasi *text mining*. Metode melakukan pengolahan teks yang dapat digunakan dalam pengklasifikasian dokumen cukup banyak, diantaranya *cosine similarity*.

Metode perhitungan TF-IDF dan algoritma *Cosine Similarity* yang digunakan di dalam penelitian ini. *Cosine Similarity* merupakan algoritma yang akan digunakan untuk menghitung kemiripan dokumen sejumlah n serta memiliki pengaruh besar terhadap nilai *similarity*. Metode yang digunakan untuk proses perhitungan nilai kesamaan atau similaritas dari dua dokumen berbeda [3].

Adanya tujuan dalam penelitian yang dilakukan adalah untuk membuat aplikasi yang dapat melakukan mengategorikan dokumen dengan otomatis dalam proses pengklasterannya penelitian ini menggunakan *cosine similarity* dan perhitunggan TF-IDF. serta mengimplementasikan rancangan model yang dibuat ke dalam aplikasi berbasis web.

2 Metode Penelitian

Untuk membangun sistem pengklasifikasian kategori topik ini menggunakan metode perhitungan TF-IDF terhadap algoritma *cosine similarity*, terdapat beberapa tahapan yang menjadi rancangan utama, rancangan ini sebagai gambaran proses tahapan awal hingga akhir sistem berjalan. Pada Gambar 1 merupakan gambar berisi Tahapan Metode.



Gambar. 1. Tahapan Penelitian

Berikut merupakan tahapan *preprocessing* :

1. *Case Folding*
   Case Folding adalah proses untuk mengubah semua karakter yang dari huruf besar (*uppercase*) menjadi huruf kecil (*lowercase*) [4] .
2. *Stemming*

Stemming merupakan proses untuk dapat meneumkan kata dasar dari sebuah teks dengan cara menghapus atau menghilangkan imbuhan yang ada yakni awalan, akhiran, sisipan serta gabungan dari awalan dan akhiran[5].

3. *Stopwords Removal*
   Stop word removal adalah proses untuk menghapus kata yang tidak sesuai pada hasil parsing dokumen berupa teks yakni dengan membandingkan kata tersebut dengan *stoplist* [6].
4. *Tokenizing*
   Tokenizing merupakan proses memisahkan kalimat menjadi penggalan kata berupa token sebelum dianalisis lebih lanjut [7].

Setelah melalui tahap preprocessing, dokumen yang digunakan sebagai data latih akan disimpan ke dalam database, namun sifatnya sementara, maksudnya adalahketika data latih yang baru di-*import* maka data latih sebelumnya akan otomatis tergantikan oleh data latih yang baru, sedangkan dokumen data uji yang sudah diinput akan tetap disimpan ke dalam *database*. Dokumen yang digunakan adalah dokumen bertipe *.xls (excel)*, dan dokumen tersebut berupa text. Selanjutnya dokumen yang telah melalui tahapan preprocessing, dilakukan perhitungan menggunakan pembobotan TF-IDF, kemudian setelah hasilnya didapatkan, maka proses selanjutnya yaitu pencarian nilai similarity dari sebuah dokumen judul menggunakan metode *cosine similarity*. Selanjutnya admin atau petugas bisa melihat proses perhitungan dengan pembobotan TF-IDF, nilai similarity serta hasil pengklasifikasian dari dokumen abstrak skripsi mahasiswa TI Universitas Budi Luhur masuk ke dalam kategori topik apa.

Dalam proses perhitungan TF-IDF menggambarkan seberapa pentingnya kata demi kata dalam sebuah dokumen [8]. Perhitungan ini digunakan untuk mendapatkan nilai bobot relevansi *term* dari sebuah dokumen terhadap keseluruhan data. Dalam TF-IDF perhitungan mengenai nilai statistik dapat mengevaluasi pentingnya kata dalam dokumen [9]. Metode pada perhitungan TF-IDF ini merupakan gabungan dari dua konsepsi menghitung nilai atau bobot, yakni frekuensi pada munculnya teks berupa kata pada dokumen tertentu dan *inverse* frekuensi dokumen yang mengandung kata tersebut di dalamnya. Banyaknya kemunculan teks atau kata sebuah dokumen menunjukkan betapa pentingnya kata pada dokumen tersebut. Berikut adalah rumus perhitungan TF-IDF :

$$Wdt = tfdt \times Idft \qquad (1)$$

Keterangan Variabel Rumus :
Wdt = nilai bobot dokumen ke-d terhadap kata ke-t
tfdt = banyaknya jumlah kata yang dicari dalam sebuah dokumen
Idft = Inversed Document Frequency (log (N/df))
N = jumlah keseluruhan dokumen
df = banyaknya jumlah dokumen yang mengandung kata yang dicari

Cosine Similarity merupakan salah satu metode yang digunakan untuk menghitung nilai similaritas antara dua dokumen. Cara atau metode ini berfokus pada proses perhitungan nilai *similarity* di antara dua vektor di dalam suatu dimensi ruang yang didapatkan dari nilai sudut cos dari perkalian dua vektor yang akan dibandingkan, dikarenakan sudut cos dari 0 bernilai 1 maka untuk nilai sudut yang lain akan bernilai kurang dari 1 [10]. Berikut merupakan rumus *consine similarity* :

$$Cos\ \alpha = \frac{N \bullet S}{|N||S|} = \frac{\sum_{i=1}^{n} N_i \times S_i}{\sqrt{\sum_{i=1}^{n}(N_i)^2} \times \sqrt{\sum_{i=1}^{n}(S_i)^2}} \qquad (2)$$

Keterangan Variabel Rumus :
N = Vektor N, ini merupakan variable yang akan dibandingkan similaritasnya
S = Vektor S, ini merupakan variable yang akan dibandingkan similaritasnya
N • S = *dot product* di antara vektor N serta vektor S
|N| = panjang vektor N
|S| = panjang vektor S
|N||S| = *cross product* antara variabel |N| dan |S|

Pada sistem ini cara atau metode yang digunakan dalam melakukan pengklasifikasian yaitu dengan membandingkan tingkat nilai *similarity* terhadap judul skripsi di data latih dan dengan judul skripsi pada data yang akan dilakukan pengujian. Kemudian dihitung dan dicari nilai *similarity* tertinggi.

3 Hasil Dan Pembahasan

3.1 Rancangan Pengujian

Untuk dapat mengetahui nilai similaritas dari sebuah dokumen pada penelitian ini maka diperlukannya sistem pengujian yakni dengan pengujian model. Pengujian model merupakan pengujian yang menjelaskan mengenai bagaimana proses sistem yang dibuat dapat berjalan, yang dirincikan prosesnya yakni dimulai dari pengunggahan dokumen, kemudian menetapkan dokumen skripsi sebagai data latih serta data uji sampai dapat menentukan klasifikasi.

3.2 Topik Penelitian pada Data Dokumen

Data dokumen yang digunakan memiliki 15 kelas atau label berupa jenis topik pada program studi Teknik Informatika Universitas Budi Luhur. Berikut Tabel 1 berisi list topik atau kategori skripsi sebagai berikut :

Tabel 1. Penggunaan Kelas (Label)

No	Topik dalam Penelitian Skripsi
1.	Kriptografi
2.	Steganografi
3.	Sistem Pakar
4.	Sistem Penunjang Keputusan (SPK)
5.	Data Mining
6.	Text Mining
7.	Natural Language Processing
8.	Pengolahan Citra Digital
9.	Otomasi Berbasis Sensor
10.	Security
11.	Augmented Reality
12.	Game Development
13.	WEB Service/API
14.	Sistem Kendali Berbasis Internet of Things
15.	Jaringan Syaraf Tiruan (JST)

3.3 Data Latih

Data latih merupakan tahap untuk pengumpulan dataset yang dilakukan dengan cara mengimport satu file bertipe file excel (.xls) yang berisi 300 data. Dokumen berisi judul tugas akhir mahasiswa Teknik Informatika Universitas Budi Luhur dalam bentuk excel(.xls) merupakan data yang digunakan dalam penelitian ini yang kemudian akan disimpan ke dalam database. Data dokumen didapatkan dari perpustakaan Universitas Budi Luhur. Pembagian untuk data latih dan data uji adalah dengan metode *Hold Out*, metode ini merupakan cara untuk melakukan pembagian data keseluruhan (dataset) menjadi beberapa bagian, untuk pembagian bagian akurasi data latih dengan data uji menggunakan persentase 90/10, maksudnya yakni 90% sebagai data latih dan 10% untuk data uji.

Sebelumnya data dokumen tersebut dilakukan pelabelan secara manual oleh penulis yang kemudian dilakukan pelabelan melalui pakar yang berjumlah satu orang, pakar tersebut adalah Bapak Dr. Indra, S.Kom., M.T.I sebagai Ketua Program Studi Teknik Informatika Universitas Budi Luhur, dikarenakan data untuk penelitian yang digunakan merupakan data skripsi mahasiswa dari jurusan Teknik Informatika maka yang dapat melakukan validasi pelabelan adalah beliau sebagai kaprodi dengan cara melakukan verifikasi dari pelabelan sebelumnya sehingga proses pelabelan tersebut menjadi valid. Data dokumen yang sudah berbentuk teks akan dilakukan tahap *preprocessing*, perhitungan bobot nilai dengan TF-IDF, serta proses perhitungan dengan algoritma *Cosine Similarity*. Sehingga diketahui tingkat similarity dari sebuah dokumen yang telah dilakukan pelabelan sebelumnya. Pada Tabel 2 merupakan contoh data dokumen yang sebelumnya belum dilakukan pelabelan.

Tabel 2. Dokumen Judul Skripsi dari Perpustakaan Universitas Budi Luhur

No	Data Dokumen
1.	Judul : IOT SENSOR GERAK PIR DAN ULTRASONIK UNTUK MEMBUKA PINTU PALANG PARKIR DAN MENGIDENTIKASI JENIS KENDARAAN YANG MASUK PARKIR DI RUKO SALEMBA MAS Nama : Afga Syaheta Pradana NIM : 1211530108

2.	Judul : KRIPTOGRAFI PENGAMANAN REST API MENGGUNAKAN METODE RSA DAN VIGENERE CIPHER UNTUK KEGIATAN MONITORING BOT PENJUALAN B2B DENGAN APLIKASI ANDROID PADA PT. GREAT GIANT LIVESTOCK Nama : Andika Ferdiyan Syah NIM : 1311500092

3.4 Hasil Pengujian Klasifikasi dengan Algoritma *Cosine Similarity*

Dalam melakukan pengujian aplikasi maka terdapat beberapa fungsi dilakukannya pengujian yakni mengevaluasi, mengetahui serta menganalisa nilai akurasi ataupun tingkat kesamaan dari dibuatnya sistem tersebut. Data yang digunakan untuk melakukan pengujian sebesar 10% dari total jumlah data yang ada yakni 300 data. Sehingga jumlah data pengujian sebanyak 30 data. Tabel 3 berisi *sample* hasil pengujian.

Tabel 3. Hasil Pengujian Algoritme

No	Judul Tugas Akhir	Klasifikasi *Actual*	Hasil klasifikasi	Hasil
1	IMPLEMENTASI DATA MINING KLASIFIKASI UNTUK MEMPREDIKSI KELULUSAN UJIAN NASIONAL SISWA SMP NEGERI 6 KOTA TANGERANG DENGAN METODE K-NEAREST NEIGHBOR BERBASIS JAVA	Data Mining	Data Mining	Sesuai
2	RANCANG BANGUN ALAT PELACAK LOKASI KENDARAAN BERMOTOR MENGGUNAKAN MIKROKONTROLLER ARDUINO UNO DAN BERBASIS GPS	Otomasi Berbasis Sensor	Sistem Kendali Berbasis Internet of Things	Tidak Sesuai
3	SISTEM MONITORING DISPENSER SABUN OTOMATIS TANPA SENTUH BERBASIS ARDUINO	Otomasi Berbasis Sensor	Otomasi Berbasis Sensor	Sesuai
4	PROTOTIPE MONITORING HUJAN DAN KETINGGIAN AIR DI PINTU AIR KAPUK MENGGUNAKAN MICROCONTROLLER WEMOS D1 R1 DENGAN SENSOR ULTRASONIC HC-SR04 DI KELURAHAN KAPUK	Otomasi Berbasis Sensor	Otomasi Berbasis Sensor	Sesuai
5	IMPLEMENTASI METODE PROGRAMMABLE LOGIC CONTROLLER (PLC) PADA SISTEM KEAMANAN DENGAN MENGGUNAKAN ARDUINO UNO R3, MODULE WIFI ESP8266, SENSOR PIR, DAN SENSOR RFID-	Otomasi Berbasis Sensor	Otomasi Berbasis Sensor	Sesuai

	RC522 BERBASIS WEB DAN E-MAIL PADA PT BETA MEDICAL			

Berdasarkan hasil pengujian diatas, maka pada Tabel 4 berikut merupakan hasil pengujian akurasi.

Tabel 4. Hasil Jumlah Pengujian Data

Terklasifikasi Benar	Terklasifikasi Salah	Total
26	4	30

3.5 Pengujian Akurasi

Pengujian yang dilakukan menggunakan uji kepakaran. Sehingga rumus pengujian akurasi adalah jumlah terklasifikasi benar / total data x 100%. Untuk nilai akurasi program tidak bernilai konstan dikarenakan penggunaan pembagian data dengan *Hold Out* menjadikan data diuji secara acak setiap dilakukan.

$$Akurasi = \frac{Jumlah\ Terklasifikasi\ Benar}{Total\ Data} \times 100\%$$

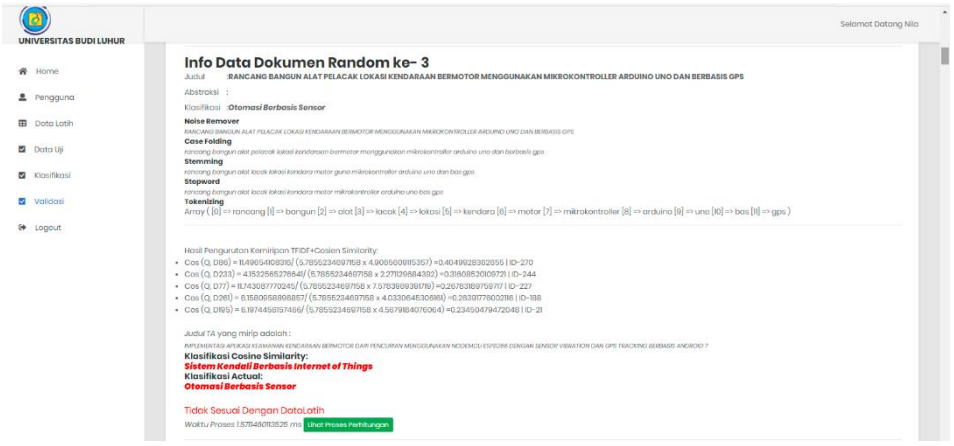
$$Akurasi = \frac{26}{30} \times 100\%$$

$$Akurasi = 86{,}66\ \%$$

Hasil uji coba dari pencarian nilai *similarity* menggunakan perhitungan TF-IDF dan algoritma *Cosine Similarity* pada data Judul Tugas Akhir Mahasiswa Teknik Informatika Universitas Budi Luhur menjelaskan pengaruh setiap term terhadap hasil *similarity*. Hasil pengujian terhadap pembagian data sebesar 90/10, maka 10% data dari 300 data yang ada menunjukan hasil akurasi sebesar 86,66%.

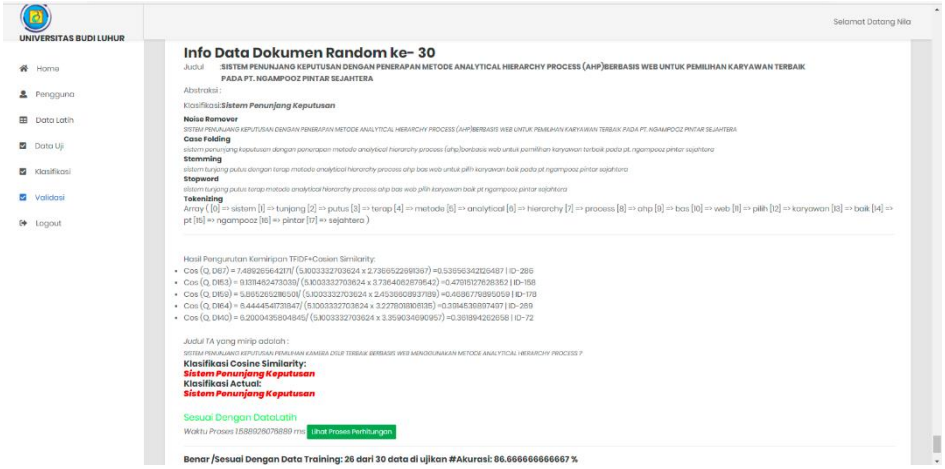
3.6 Implementasi Rancangan Sistem

Prosedur Sistem Klasifikasi Topik Tugas Akhir ini terdapat dua *level user*, yakni yang pertama *level* admin dan kedua *level* pengguna. Halaman utama (Home) pada admin terdiri dari beberapa menu diantaranya menu Pengguna, Data Latih, Data Uji, Klasifikasi, Validasi, dan Logout. Hak untuk mengakses pada admin adalah segala aspek yang terdapat dalam sistem, seperti melakukan unggah (*import*) file data latih bertipe .xls (excel), menambahkan,mengubah, serta menghapus data pengguna lain, melakukan pengujian, melihat hasil pengujian. Halaman pengguna terdiri dari beberapa menu diantaranya adalah menu Profil, Data Latih, Data Uji, Klasifikasi, Validasi, dan Logout. Hak akses pengguna adalah mengubah data pribadi pengguna, melihat data yang telah diimport oleh admin pada menu Data Latih, melakukan proses pengujian, dan dapat melihat data hasil pengujian. Gambar 2 berikut merupakan tampilan hasil pengujian dengan hasil klasifikasi tidak sesuai dengan data *actual*.



Gambar. 2. Tampilan Hasil Klasifikasi Tidak Sesuai

Gambar 3 berikut merupakan tampilan halaman bila melihat hasil klasifikasi yang sesuai dengan data *actual* serta nilai akurasi sistem pada menu validasi.



Gambar. 3. Tampilan Bila Klasifikasi Sesuai serta Nilai Akurasi

3.7 Analisis Hasil

Proses pengujian tingkat kelayakan pada Sistem Klasifikasi Topik Tugas Akhir Mahasiswa Teknik Informatika telah dilakukan. Pengujian yang dilakukan tersebut antara lain adalah pengujian *model*. Hal yang dilakukan dalam pengujian model yakni dengan melakukan observasi, hasil akhir yang akan didapat melalui beberapa sample data kemudian melakukan pemeriksaan terhadap software. Pengujian model juga menjelaskan mengenai bagaimana proses sistem yang dibuat dapat berjalan, yang dirincikan prosesnya yakni dimulai dari pengunggahan dokumen, kemudian menetapkan dokumen skripsi sebagai data latih serta data uji sampai dapat menentukan klasifikasi. Hasil pengujian model tersebut menunjukkan bahwa aplikasi ini dalam menjalankan semua menu atau fungsinya dengan baik sehingga tidak diperlukannya perbaikan.

Dalam 30 data yang sudah dilakukan pengujian secara otomatis dengan sistem, maka terdapat 26 data yang telah berhasil dikategorikan (diklasifikasikan) dengan sesuai dan untuk hasil data tidak sesuai sebanyak 4 data. Jadi nilai keakuratan proses klasifikasi aplikasi ini sebesar 86,66%. Nilai persentase tersebut diperoleh dari banyaknya jumlah data terklasifikasi benar dibagi dengan banyaknya keseluruhan data yang dilakukan pengujian lalu perhitungan selanjutnya dikali dengan nilai 100%. Beberapa kata yang sama dengan key word menjadi salah satu penyebab kesalahan dalam proses pengklasifikasian, sehingga pada hasil perhitungan algoritma cosine similarity sistem akan memilih nilai yang paling tinggi. Dikarenakan banyaknya data latih yang ada pada sistem, sehingga jumlah kata atau term juga meningkat serta masih terdapat kata yang kurang tepat pada saat proses mencari kata dasar (*stemming*) pada suatu term.

Program ini merupakan salah satu hal yang dibutuhkan dalam setiap pengklasifikasian terhadap data terkait topik yang dibahas. Dengan proses yang terkomputerisasi dalam hal tersebut membuat pengguna yang bersangkutan tidak lagi diharuskan untuk memahami keseluruhan isi pada skripsi mahasiswa Teknik Informatika sehingga proses penetuan topik pada skripsi mahasiswa Teknik Informatika Universitas Budi Luhur lebih cepat, serta program dapat diaplikasikan tanpa adanya jaringan internet karena dijalankan dengan mengandalkan *localhost.*

4 Kesimpulan

Setelah melakukan proses Analisa, mengumpulkan data serta pengujian kelayakan sistem dapat disimpulkan bahwa Sistem Klasifikasi Topik Tugas Akhir berdasarkan dokumen judul ini telah dapat mendeteksi topik sebuah dokumen dengan sumber data dokumen skripsi mahasiswa Teknik Informatika Universitas Budi Luhur berdasarkan judul skripsi. Berdasarkan pengujian yang dilakukan oleh sistem, ada 26 dari 30 data berhasil dikategorikan (diklasifikasikan) ke dalam kategori topik yang sesuai dengan pelabelan, namun terdapat 4 dokumen yang tidak dapat dikategorikan (terklasifikasikan) dengan benar (tidak sesuai). Sehingga berdasarkan pengujian

menunjukkan hasil persentase sebesar 86,66%, maka dapat ditarik kesimpulan bahwa sistem dapat melakukan proses pengklasifikasian dokumen skripsi mahasiswa berdasarkan judul dengan nilai akurasi tinggi.

Referensi

[1] N. Suri, "Bab II Landasan Teori," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2019.
[2] A. Nurhadi, "Implementasi Algoritma Naïve Bayes Classifier Berbasis Particle Swarm Optimization (PSO) Untuk Klasifikasi Konten Berita Digital Berbahasa Indonesia," *J. Speed*, vol. 5, no. 3, pp. 7–12, 2016, [Online]. Available: http://ejurnal.net/portal/index.php/speed/article/view/799/730.
[3] E. L. Amalia, A. J. Jumadi, I. A. Mashudi, and D. W. Wibowo, "Analisis Metode Cosine Similarity Pada Aplikasi Ujian Online Otomatis (Studi Kasus JTI POLINEMA)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 2, p. 343, 2021, doi: 10.25126/jtiik.2021824356.
[4] A. Salam, J. Zeniarja, and R. S. U. Khasanah, "Analisis Sentimen Data Komentar Sosial Media Facebook Dengan K-Nearest Neighbor (Studi Kasus Pada Akun Jasa Ekspedisi Barang J&T Ekpress Indonesia)," *Pros. SINTAK*, pp. 480–486, 2018.
[5] S. Sugiyamto, B. Surarso, and A. Sugiharto, "Analisa Performa Metode Cosine Dan Jacard Pada Pengujian Kesamaan Dokumen," *J. Masy. Inform.*, vol. 5, no. 10, pp. 1–8, 2014, doi: 10.14710/jmasif.5.10.1-8.
[6] S. S. Mohammad Khoiron and Adhi Dharma Wibawa, *Microblogging Analysis for Determining Public Policy Priority based on Public Opinion using Naïve Bayes and Analytical Hierarchy Process Algorithm*, vol. 5, no. 1. 2016.
[7] H. F. Fadli and A. F. Hidayatullah, "Identifikasi Cyberbullying pada Media Sosial Twitter Menggunakan Metode LSTM dan BiLSTM," *Automata*, 2021, [Online]. Available: https://journal.uii.ac.id/AUTOMATA/article/view/17364.
[8] L. W. Astuti, A. Rachmat C., and Y. Lukito, "Implementasi Algoritma Naïve Bayes Menggunakan Isear Untuk Klasifikasi Emosi Lirik Lagu Berbahasa Inggris," *J. Inform.*, vol. 14, no. 1, pp. 16–21, 2017, doi: 10.9744/informatika.14.1.16-21.
[9] A. Afriza and J. Adisantoso, "Metode Klasifikasi Rocchio untuk Analisis Hoax Rocchio Classification Method for Hoax Analysis," *J. Ilmu Komput. Agri-Informatika*, vol. 5, pp. 1–10, 2018, [Online]. Available: http://journal.ipb.ac.id/index.php/jika.
[10] M. D. R. Wahyudi, "Penerapan Algoritma Cosine Similarity pada Text Mining Terjemah Al-Qur'an Berdasarkan Keterkaitan Topik," *Semesta Tek.*, vol. 22, no. 1, pp. 41–50, 2019, doi: 10.18196/st.221235.