Indonesian Journal of Computer Science
ISSN 2302-4364 (*print*) dan 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam, No. 1, Padang, Indonesia, Telp. (0751) 7056199, 7058325
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Penerapan Text Mining Pada Sistem Penyeleksian Judul Skripsi Menggunakan Algoritma Latent Dirichlet Allocation (LDA)

Rakhmat Kurniawan R, Ilka Zufria
rakhmat.kr@uinsu.ac.id, ilkazufria@uinsu.ac.id
Universitas Islam Negeri Sumatera Utara Medan

Informasi Artikel

Diterima : 27 Nov 2022
Direview : 22 Des 2022
Disetujui : 30 Des 2022

Kata Kunci

Skripsi, Judul, Text Mining, LDA

Abstrak

Dalam proses pengajuan judul proposal skripsi, banyak mahasiswa yang judulnya ditolak dikarenakan adanya kesamaan judul atau tema penelitian yang telah dilakukan sebelumnya. Proses ini dilakukan secara manual dimana judul skripsi direkapitulasi dengan menggunakan aplikasi Microsoft Excel, sehingga membuka peluang kekeliruan dalam pemeriksaan yang disebabkan oleh tim penyeleksi memeriksa judul secara manual baris per baris. Selanjutnya dirasa perlu melakukan penelitian untuk mengatasi masalah tersebut. Dengan harapan memudahkan pengelola program studi dalam menentukan judul skripsi yang potensial dan berkualitas pada mahasiswa. Selanjutnya mempercepat proses penentuan judul skripsi mahasiswa dan tentunya juga akan mempercepat proses penyelesaian studi mahasiswa strata 1. Maka dirancanglah sebuah sistem yang mampu merekomendasi kelayakan judul skripsi yang diajukan melalui teknologi text mining menggunakan algoritma LDA. Algoritma LDA mampu untuk mendeteksi topik yang ada pada suatu koleksi dokumen beserta besarnya kemunculan topik tersebut.

Keywords

Thesis, Title, Text Mining, LDA

Abstrak

In the process of submitting the title of the thesis proposal, many students whose titles were rejected due to the similarity of titles or research themes that had been carried out previously. This process is carried out manually where the thesis title is recapitulated using the Microsoft Excel application, thus opening up opportunities for errors in the examination caused by the selection team checking the title manually line by line. Furthermore, it is deemed necessary to conduct research to overcome these problems. With the hope of making it easier for study program managers in determining potential and quality thesis titles for students. Furthermore, accelerating the process of determining the title of student thesis and of course will also speed up the process of completing the study of undergraduate students. So a system is designed that is able to recommend the feasibility of the proposed thesis title through text mining technology using the LDA algorithm. The LDA algorithm is able to detect the topics that exist in a collection of documents and the magnitude of the occurrence of these topics.

A. Pendahuluan

Skripsi merupakan karya tulis ilmiah yang disusun oleh mahasiswa pada tingkat Strata I yang dilaksanakan melalui field research ataupun library research. Penelitian ini dilakukan untuk memenuhi salah syarat untuk memperoleh Gelar Sarjana. Setiap mahasiswa yang telah memenuhi syarat, diwajibkan untuk mengajukan judul skripsi. Setiap judul skripsi yang telah diajukan ke program studi, akan melalui proses seleksi yang ketat oleh tim penyeleksi. Banyak mahasiswa yang judulnya tertolak dikarenakan kesamaan judul atau tema penelitian yang telah dilakukan sebelumnya. Proses ini dilakukan secara manual dimana judul skripsi direkapitulasi dengan menggunakan aplikasi Microsoft Excel, sehingga membuka peluang kekeliruan dalam pemeriksaan yang disebabkan oleh tim penyeleksi memeriksa judul secara manual baris per baris. Saat ini di Program Studi Ilmu Komputer menerima pengajuan proposal judul skripsi mahasiswa lebih dari 500 judul per semester. Selain kesamaan judul dan tema, kontribusi penelitian terhadap ilmu pengetahuan khususnya Ilmu Komputer juga ikut menentukan diterima atau ditolaknya judul skripsi.

Text Mining merupakan suatu proses eksplorasi dan analisis dataset yang besar dalam bentuk text untuk mendapatkan suatu informasi yang bermanfaat untuk tujuan tertentu. Salah satu algoritma yang dapat digunakan dalam text mining adalah LDA (Latent Dirichlet Allocation). Algoritma LDA mampu untuk mendeteksi topik yang ada pada suatu koleksi dokumen beserta besarnya kemunculan topik tersebut pada koleksi dokumen maupun di dokumen tertentu.

Pada penelitian terdahulu yang dilakukan oleh Ayu Lestari pada tahun 2018 dengan judul "Implementasi Data Mining Untuk Seleksi Judul Skripsi Mahasiswa Menggunakan Metode Association Rule (Studi Kasus: Fakultas Ilmu Komputer Universitas Pasir Pengaraian)" hanya memberikan output judul diterima atau ditolak, sedangkan dalam penelitian ini output yang dihasilkan tidak hanya diterima atau ditolaknya judul skripsi tersebut, namun akan memberikan rekomendasi topik yang potensial untuk diterima.

Berdasarkan uraian diatas, maka Program Studi Ilmu Komputer dirasa perlu untuk melakukan penelitian terkait hal tersebut. Dengan demikian memudahkan pengelola program studi dalam menentukan judul skripsi yang potensial dan berkualitas. Selanjutnya mempercepat proses penentuan judul skripsi mahasiswa dan tentunya juga akan mempercepat proses penyelesaian studi strata I di program studi ilmu komputer UIN Sumatera Utara Medan dan meningkatkan jumlah mahasiswa yang lulus tepat waktu.

Berdasarkan permasalahan dan uraian penyelesaian solusinya, maka diciptakan penelitian dengan judul: "Penerapan Text Mining Pada Sistem Penyeleksian Judul Skripsi Mahasiswa Menggunakan Algoritma LDA (Latent Dirichlet Allocation) di Program Studi Ilmu Komputer UIN Sumatera Utara Medan".

B. Metode Penelitian

1. Jenis Metode

Jenis penelitian yang diadopsi dalam penelitian ini adalah penelitian *R&D (Research and Development)* dan menggunakan metode waterfall untuk pengembangan sistem. Pemilihan metode *R&D* berdasarkan pada tujuan penelitian unruk menghasilkan produk yang efisien. Metode penelitian *Research and Development*, memiliki beberapa tahapan sebagaimana dikemukakan oleh (Sugiyono, 2016) yaitu: "1. Potensi dan masalah, 2. Pengumpulan data, 3. Desain produk, 4. Validasi Desain, 5. Perbaikan Desain, 6. Uji Coba Produk, 7. Revisi Produk, 8. Uji Pelaksanaan Lapangan, 9.

Penyempurnaan Produk, 10. Implementasi ".

2. Pendekatan Penelitian

Penelitian ini menggunakan tipe penelitian eksperimental. Menurut (Sugiyono, 2016) penelitian eksperimental adalah penelitian eksperimen yaitu penelitian yang digunakan untuk mencari pengaruh perlakuan tertentu terhadap yang lain dalam kondisi yang terkendali. (Arikunto, 2017) juga berpedapat yang sama, mendefinisikan penelitian eksperimen merupakan penelitian yang dimaksudkan untuk mengetahui ada tidaknya akibat dari treatment pada subjek yang diselidiki. Cara untuk mengetahuinya yaitu membandingkan satu atau lebih kelompok eksperimen yang diberi treatment dengan satu kelompok pembanding yang tidak diberi treatment.

Menurut (Sugiyono, 2016) terdapat beberapa bentuk desain eksperimen yaitu:

a. Pre-experimental design
b. True experimental design
c. Factorial design
d. *Quasi experimental design*.

(Sugiyono, 2016) juga menyatakan bahwa ciri utama dari quasi experimental design adalah pengembangan dari true experimental design, yang mempunyai kelompok control namun tidak dapat berfungsi sepenuhnya untuk mengontrol variabel-variabel dari luar yang mempengaruhi eksperimen.

Berdasarkan uraian diatas dapat diambil kesimpulan bahwa *quasi experimental design* adalah jenis penelitian yang memiliki kelompok kontrol dan kelompok eksperimen tidak dipilih secara acak. Pada penelitian ini peneliti menggunakan quasi experimental design karena dalam penelitian ini terdapat variabel-variabel dari luar yang tidak dapat dikontrol oleh peneliti.

3. Pengumpulan Data

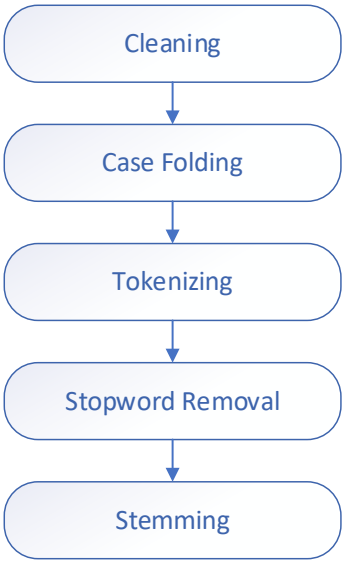
Data yang digunakan dalam penelitian ini berupa dokumen proposal skripsi mahasiswa Program Studi Ilmu Komputer UIN Sumatera Utara Medan yang diperoleh langsung dari Program Studi Ilmu Komputer. Data yang digunakan adalah proposal skripsi Program Studi Ilmu Komputer dari tahun 2020 – 2022 dengan jumlah 395 rows. Data ini akan terus bertambah seiring berjalannya waktu. Data yang diperoleh dari Program Studi Ilmu Komputer berupa file spreadsheet dalam format gsheet. File tersebut berisi field-field judul, dimana tiap row terdiri dari 3 (tiga) judul proposal. Ketiga judul proposal pada setiap row ini nantinya akan dipecah menjadi 3 (tiga) row. Dengan demikian data asli yang berjumlah 395 row nantinya akan pecah hingga menjadi 1185 row.

4. Analisis Data

Dalam melakukan analisa data yang akan melalui 3 (tiga) tahap pemrosesan, yaitu *preprocesseing* data, pembobotan kata dengan TF-IDF, dan *topic modeling* dengan menggunakan *Latent Dirichlet Allocation(LDA)*. Pada tahap *preprocessing* data, dilakukan pembersihan data berupa tanda baca dan *stopword* yang sesuai dengan Bahasa Indonesia yang baku. Pada tahapan berikutnya yaitu pembobotan kata yang dilakukan dengan menggunakan TF-IDF, dimana setiap kata akan dijadikan data numerik dari dokumen yang akan diproses. Tahap akhir dari analisa data *adalah topic modelling* dimana algoritma *Latent Dirichlet Allocation* yang akan mengolah data hasil pembobotan menjadi hasil dari pemodelan topik.

a. *Preprocessing* Data

Langkah awal dalam penelitian ini sebelum melakukan tahap pengelompokan teks yaitu proses preprocessing data. Tahap preprocessing data ini dilakukan untuk membersihkan atau menghilangkan teks dalam dokumen yang tidak diperlukan. Tahapan preprocessing ini akan melalui 5 (lima) langkah proses sebelum dilakukan pembobotan pada tahap berikutnya. 5 (langkah) proses preprocessing dapat dilihat pada Gambar 1 dibawah ini.



Gambar 1. Tahapan *preprocessing*

b. Pembobotan TF-IDF

Proses TF-IDF dilakukan untuk memperoleh nilai numeric dari setiap teks yang akan dikelompokkan. Tahap TF – IDF dimulai dari menghitung nilai Tf yang merupakan nilai teks berdasarkan seberapa sering teks tersebut muncul pada data, dilanjutkan dengan proses perhitungan idf yang berfokus pada seberapa sering teks muncul pada data yang berbeda dan diakhiri dengan proses perkalian antara nilai tf dan idf untuk memproleh nilai numeric dari setiap teks, sehingga dapat dilakukan pengelompokkan dengan algoritma *Latent Dirichlet Allocation*.

c. Topic Modelling

Proses *topic modeling* menggunakan algoritma *Latent Dirichlet Allocation* dimulai dengan merubah kata yang terdapat pada teks judul skripsi kedalam bentuk numeric dan melakukan perhitungan frekuensi penyebaran kata menggunakan metode TF – IDF, dilanjukan dengan menentukan jumlah ttopic yang digunakan, dalam penelitian ini menggunakan lebih dari satu topic. Untuk mengetahui teks judul skripsi yang digunakan masuk kedalam topic 1, topic 2 atau topic n dapat dilakukan perhitungan probabilitas, proses penentuan topic dari teks berdasarkan nilai probabilitas terbesar dari beberapa topic yang telah ditentukan.

C. Hasil dan Pembahasan

a. Representasi Data

Data yang digunakan dalam penelitian ini merupakan judul skripsi yang diajukan mahasiswa kepada program studi Ilmu Komputer Universitas

Islam Negeri Sumatera Utara dalam priode pengajuan tahun 2019 – 2022 dengan jumlah 395 row. Pada masing-masing row memiliki tiga kolom judul, sehingga total keseluruhan dari judul skripsi yang digunakan pada penelitian inii sebanyak 1185 row.

b. Hasil Analisis Data

Hasil analisis data digunakan untuk mengetahui apakah data yang digunakan sesuai dengan penelitian. Tahap awal adalah preprocessing teks yang bertujuan untuk merubah teks menjadi data terstruktur. Berikut ini merupakan sample teks berupa judul yang digunakan.

Tabel 1. Data Sampel Judul Skripsi

Judul Skripsi
IMPLEMENTASI METODE PREFERENCE RANKING ORGANIZATION METHOD FOR ENRICHMENT EVALUATION (PROMETHEE II) UNTUK PENENTUAN KENAIKAN JABATAN KARYAWAN PT AGUNG SARANA UTAMA
Implementasi Algoritma Principal Component Analysis (PCA) Untuk Aplikasi Absensi Pengenalan Wajah
Implementasi Data Mining Menggunakan Algoritma C4.5 untuk Penentuan Penerima Bantuan Siswa Miskin Pada Masa Pandemi Covid-19 di SMKN 1 Lubuk Pakam
Implementasi Pengamanan File Citra Bitmap dengan Algoritma Chacha20
analisis perbandingan penentuan jalur terpendek dengan menggunakan algoritma general and test dengan best first search

1) Cleaning

Cleaning dilakukan untuk membersihkan data (teks) dari noise berupa tanda baca. Berikut ini teks hasil cleaning

Tabel 2. Hasil *Cleaning*

Judul Skripsi
IMPLEMENTASI METODE PREFERENCE RANKING ORGANIZATION METHOD FOR ENRICHMENT EVALUATION PROMETHEE II UNTUK PENENTUAN KENAIKAN JABATAN KARYAWAN PT AGUNG SARANA UTAMA
Implementasi Algoritma Principal Component Analysis PCA Untuk Aplikasi Absensi Pengenalan Wajah
Implementasi Data Mining Menggunakan Algoritma C4 5 untuk Penentuan Penerima Bantuan Siswa Miskin Pada Masa Pandemi Covid 19 di SMKN 1 Lubuk Pakam
Implementasi Pengamanan File Citra Bitmap dengan Algoritma Chacha20
analisis perbandingan penentuan jalur terpendek dengan menggunakan algoritma general and test dengan best first search

2) Case Folding

Tahapan ini bertujuan untuk melakukan penyeragaman case menjadi huruf kecil pada penulisan judul skripsi, dikarenakan bahasa yang digunakan dalam penerapan topic modeling merupakan case sensitive yang menganggap judul skripsi yang ditulis dengan huruf kapital dan huruf kecil berbeda walaupun sebenarnya kedua judul tersebut sama dan hal ini dapat menyebabkan terjadinya data kembar.

Tabel 3. Tabel Hasil *Case Folding*

Judul Skripsi
implementasi metode preference ranking organization method for enrichment evaluation promethee ii untuk penentuan kenaikan jabatan karyawan pt agung

sarana utama
implementasi algoritma principal component analysis pca untuk aplikasi absensi pengenalan wajah
implementasi data mining menggunakan algoritma c4 5 untuk penentuan penerima bantuan siswa miskin pada masa pandemi covid 19 di smkn 1 lubuk pakam
implementasi pengamanan file citra bitmap dengan algoritma chacha20
analisis perbandingan penentuan jalur terpendek dengan menggunakan algoritma general and test dengan best first search

3) Tokenizing

Tahapan tokenizing untuk memecah teks judul skripsi menjadi elemen kecil berupa kata tunggal yang dipisahkan dengan tanda petik (‘ ‘) sehingga dapat mempermudah tahapn selanjutnya. Berikut ini teks hasil tokenizing

Tabel 4. Tabel Hasil *Tokenizing*

Judul Skripsi
‘implementasi’, ‘metode’, ‘preference’, ‘ranking’, ‘organization’, ‘method’, ‘for’, ‘enrichment’, ‘evaluation’, ‘promethee’, ‘ii’, ‘untuk’, ‘penentuan’, ‘kenaikan’, ‘jabatan’, ‘karyawan’, ‘pt’, ‘agung’, ‘sarana’, ‘utama’
‘implementasi’, ‘algoritma’, ‘principal’, ‘component’, ‘analysis’, ‘pca’, ‘untuk’, ‘aplikasi’, ‘absensi’, ‘pengenalan’, ‘wajah’
‘implementasi’, ‘data’, ‘mining’, ‘menggunakan’, ‘algoritma’, ‘c4’, ‘5’, ‘untuk’, ‘penentuan’, ‘penerima’, ‘bantuan’, ‘siswa’, ‘miskin’, ‘pada’, ‘masa’, ‘pandemi’, ‘covid’, ‘19’, ‘di’, ‘smkn’, ‘1’, ‘lubuk’, ‘pakam’
‘implementasi’, ‘pengamanan’, ‘file’, ‘citra’, ‘bitmap’, ‘dengan’, ‘algoritma’, ‘chacha20’
‘analisis’, ‘perbandingan’, ‘penentuan’, ‘jalur’, ‘terpendek’, ‘dengan’, ‘menggunakan’, ‘algoritma’, ‘general’, ‘and’, ‘test’, ‘dengan’, ‘best’, ‘first’, ‘search’

4) Stopword Removal

Stopword removal digunakan untuk menghapus kata – kata yang tidak memiliki makna penting dalam teks judul skripsi. Berikut ini hasil *stopword removal*

Tabel 5. Tabel Hasil *Stopword Removal*

Judul Skripsi
‘implementasi’, ‘metode’, ‘preference’, ‘ranking’, ‘organization’, ‘method’, ‘for’, ‘enrichment’, ‘evaluation’, ‘promethee’, ‘ii’, ‘penentuan’, ‘kenaikan’, ‘jabatan’, ‘karyawan’, ‘pt’, ‘agung’, ‘sarana’, ‘utama’
‘implementasi’, ‘algoritma’, ‘principal’, ‘component’, ‘analysis’, ‘pca’, ‘aplikasi’, ‘absensi’, ‘pengenalan’, ‘wajah’
‘implementasi’, ‘data’, ‘mining’, ‘menggunakan’, ‘algoritma’, ‘c4’, ‘5’, , ‘penentuan’, ‘penerima’, ‘bantuan’, ‘siswa’, ‘miskin’, ‘pandemi’, ‘covid’, ‘19’, ‘smkn’, ‘1’, ‘lubuk’, ‘pakam’
‘implementasi’, ‘pengamanan’, ‘file’, ‘citra’, ‘bitmap’, ‘algoritma’, ‘chacha20’
‘analisis’, ‘perbandingan’, ‘penentuan’, ‘jalur’, ‘terpendek’, ‘menggunakan’, ‘algoritma’, ‘general’, ‘and’, ‘test’, ‘best’, ‘first’, ‘search’

5) Stemming

Tahapan untuk menghapus imbuhan yang terdapat pada kata baik itu berupa suffix, prefix maupun affix. Sehingga setiap kata akan kembali kebentuk kata dasar. Berikut ini hasil tahapan *stemming*

Tabel 6. Tabel Hasil *Stemming*

Judul Skripsi
'implementasi', 'metode', 'preference', 'ranking', 'organization', 'method', 'for', 'enrichment', 'evaluation', 'promethee', 'ii', 'tentu', 'naik', 'jabat', 'karyawan', 'pt', 'agung', 'sarana', 'utama'
'implementasi', 'algoritma', 'principal', 'component', 'analysis', 'pca', 'aplikasi', 'absensi', 'kenal', 'wajah'
'implementasi', 'data', 'mining', 'guna', 'algoritma', 'c4', '5', , 'tentu', 'terima', 'bantu', 'siswa', 'miskin', 'pandemi', 'covid', '19', 'smkn', '1', 'lubuk', 'pakam'
'implementasi', 'aman', 'file', 'citra', 'bitmap', 'algoritma', 'chacha20'
'analisis', 'banding', 'tentu', 'jalur', 'pendek', 'guna', 'algoritma', 'general', 'and', 'test', 'best', 'first', 'search'

6) **Pembobotan TF-IDF**

TF – IDF dilakukan untuk mencari nilai *numeric* dari setiap kata yang terdapat pada *teks* hasil *preprocessing* berdasarkan intensitas kemunculan kata tersebut. Tahap awal yang di lakukan adalah menghitung nilai TF dari semua kata yang terdapat dalam *teks*, berikut ini hasil perhitungan nilai TF :

Tabel 7. Tabel TF

Kata	TF					DF
	T1	T2	T3	T4	T5	
implementasi	1	1	1	1	0	4
metode	1	0	0	0	0	1
preference	1	0	0	0	0	1
ranking	1	0	0	0	0	1
organization	1	0	0	0	0	1
method	1	0	0	0	0	1
for	1	0	0	0	0	1
enrichment	1	0	0	0	0	1
evaluation	1	0	0	0	0	1
promethee	1	0	0	0	0	1
ii	1	0	0	0	0	1
tentu	1	0	1	0	1	3
naik	1	0	0	0	0	1
jabat	1	0	0	0	0	1
karyawan	1	0	0	0	0	1
pt	1	0	0	0	0	1
agung	1	0	0	0	0	1
sarana	1	0	0	0	0	1
utama	1	0	0	0	0	1
algoritma	0	1	1	1	1	4
principal	0	1	0	0	0	1
component	0	1	0	0	0	1
analysis	0	1	0	0	0	1
pca	0	1	0	0	0	1
aplikasi	0	1	0	0	0	1
absen	0	1	0	0	0	1
kenal	0	1	0	0	0	1
wajah	0	1	0	0	0	1
data	0	0	1	0	0	1
mining	0	0	1	0	0	1

c4	0	0	1	0	0	1
terima	0	0	1	0	0	1
bantu	0	0	1	0	0	1
siswa	0	0	1	0	0	1
miskin	0	0	1	0	0	1
pandemi	0	0	1	0	0	1
covid	0	0	1	0	0	1
smkn	0	0	1	0	0	1
lubuk	0	0	1	0	0	1
pakam	0	0	1	0	0	1
aman	0	0	0	1	0	1
file	0	0	0	1	0	1
citra	0	0	0	1	0	1
bitmap	0	0	0	1	0	1
analisis	0	0	0	0	1	1
banding	0	0	0	0	1	1
jalur	0	0	0	0	1	1
pendek	0	0	0	0	1	1
general	0	0	0	0	1	1
and	0	0	0	0	1	1
test	0	0	0	0	1	1
best	0	0	0	0	1	1
first	0	0	0	0	1	1
search	0	0	0	0	1	1

Tahap selanjutnya menghitung nilai IDF menggunakan persamaan $idf_t = \log \frac{N}{dft}$, berikut ini adalah proses perhitungannya,

a. $idf_{\text{implementasi}} = \log \frac{4}{5} = -0{,}97$
b. $idf_{\text{metode}} = \log \frac{1}{5} = -0{,}699$
c. $idf_{\text{preference}} = \log \frac{1}{5} = -0{,}699$
d. $idf_{\text{ranking}} = \log \frac{1}{5} = -0{,}699$
e. $idf_{\text{organization}} = \log \frac{1}{5} = -0{,}699$

Untuk proses perhitungan nilai IDF tidak dijabarkan keseluruhan proses perhitungan yang terdapat pada *sample teks* yang digunakan. Untuk hasil nilai IDF kesuluruhan kata telah disajikan dalam bentuk tabel dibawah ini

Tabel 8. Tabel IDF

Kata	DF	IDF
implementasi	4	-0,097
metode	1	-0,699
preference	1	-0,699
ranking	1	-0,699
organization	1	-0,699
method	1	-0,699
for	1	-0,699
enrichment	1	-0,699
evaluation	1	-0,699

promethee	1	-0,699
ii	1	-0,699
tentu	3	-0,125
naik	1	-0,699
jabat	1	-0,699
karyawan	1	-0,699
pt	1	-0,699
agung	1	-0,699
sarana	1	-0,699
utama	1	-0,699
algoritma	4	-0,097
principal	1	-0,699
component	1	-0,699
analysis	1	-0,699
pca	1	-0,699
aplikasi	1	-0,699
absen	1	-0,699
kenal	1	-0,699
wajah	1	-0,699
data	1	-0,699
mining	1	-0,699
c4	1	-0,699
terima	1	-0,699
bantu	1	-0,699
siswa	1	-0,699
miskin	1	-0,699
pandemi	1	-0,699
covid	1	-0,699
smkn	1	-0,699
lubuk	1	-0,699
pakam	1	-0,699
aman	1	-0,699
file	1	-0,699
citra	1	-0,699
bitmap	1	-0,699
analisis	1	-0,699
banding	1	-0,699
jalur	1	-0,699
pendek	1	-0,699
general	1	-0,699
and	1	-0,699
test	1	-0,699
best	1	-0,699
first	1	-0,699
search	1	-0,699

Setelah diperoleh nilai TF dari semua kata yang terdapat dalam *teks* judul skripsi, dilanjutkan dengan menghitung nilai TF – IDF, dengan cara melakukan perkalian antara nilai TF dan IDF dari masing – masing kata yang telah diperoleh. Berikut ini cara perhitungan TF – IDF untuk masing – masing kata.

$$tf - idf_{t.d} = tf_{td} \times idf_t$$

a. $$tf - idf\,(\text{implementasi}, \text{T1}) = tf_{\text{implementasi.T1}} \times idf_{\text{implementasi}}$$
$$= 1 \times -0{,}097 = -{,}0097$$

b. $$tf - idf\,(\text{implementasi}, \text{T2}) = tf_{\text{implementasi.T2}} \times idf_{\text{implementasi}}$$

$$= 1 \times -0{,}097 = -{,}0097$$

c. $tf - idf\,(\text{implementasi}, T3) = tf_{\text{implementasi}.T3} \times idf_{\text{implementasi}}$
$$= 1 \times -0{,}097 = -{,}0097$$

d. $tf - idf\,(\text{implementasi}, T4) = tf_{\text{implementasi}.T4} \times idf_{\text{implementasi}}$
$$= 1 \times -0{,}097 = -{,}0097$$

e. $tf - idf\,(\text{implementasi}, T5) = tf_{\text{implementasi}.T5} \times idf_{\text{implementasi}}$
$$= 0 \times -0{,}097 = 0$$

Untuk proses perhitungan nilai TF- IDF tidak dijabarkan keseluruhan proses perhitungan yang terdapat pada *sample teks* yang digunakan. Untuk hasil nilai IDF kesuluruhan kata telah disajikan dalam bentuk tabel dibawah ini

Tabel 9. Tabel TF-IDF

Kata	TF					DF	IDF	TF-IDF				
	T1	T2	T3	T4	T5			T1	T2	T3	T4	T5
implementasi	1	1	1	1	0	4	-0,097	-0,097	-0,097	-0,097	-0,097	0
metode	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
preference	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
ranking	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
organization	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
method	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
for	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
enrichment	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
evaluation	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
promethee	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
ii	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
tentu	1	0	1	0	1	3	-0,125	-0,125	0	-0,125	0	-0,125
naik	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
jabat	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
karyawan	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
pt	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
agung	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
sarana	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
utama	1	0	0	0	0	1	-0,699	-0,699	0	0	0	0
algoritma	0	1	1	1	1	4	-0,097	0	-0,097	-0,097	-0,097	-0,097
principal	0	1	0	0	0	1	-0,699	0	-0,699	0	0	0
component	0	1	0	0	0	1	-0,699	0	-0,699	0	0	0
analysis	0	1	0	0	0	1	-0,699	0	-0,699	0	0	0
pca	0	1	0	0	0	1	-0,699	0	-0,699	0	0	0
aplikasi	0	1	0	0	0	1	-0,699	0	-0,699	0	0	0
absen	0	1	0	0	0	1	-0,699	0	-0,699	0	0	0
kenal	0	1	0	0	0	1	-0,699	0	-0,699	0	0	0
wajah	0	1	0	0	0	1	-0,699	0	-0,699	0	0	0
data	0	0	1	0	0	1	-0,699	0	0	-0,699	0	0
mining	0	0	1	0	0	1	-0,699	0	0	-0,699	0	0
c4	0	0	1	0	0	1	-0,699	0	0	-0,699	0	0
terima	0	0	1	0	0	1	-0,699	0	0	-0,699	0	0
bantu	0	0	1	0	0	1	-0,699	0	0	-0,699	0	0
siswa	0	0	1	0	0	1	-0,699	0	0	-0,699	0	0
miskin	0	0	1	0	0	1	-0,699	0	0	-0,699	0	0
pandemi	0	0	1	0	0	1	-0,699	0	0	-0,699	0	0
covid	0	0	1	0	0	1	-0,699	0	0	-0,699	0	0
smkn	0	0	1	0	0	1	-0,699	0	0	-0,699	0	0

lubuk	0	0	1	0	0	1	-0,699	0	0	-0,699	0	0
pakam	0	0	1	0	0	1	-0,699	0	0	-0,699	0	0
aman	0	0	0	1	0	1	-0,699	0	0	0	-0,699	0
file	0	0	0	1	0	1	-0,699	0	0	0	-0,699	0
citra	0	0	0	1	0	1	-0,699	0	0	0	-0,699	0
bitmap	0	0	0	1	0	1	-0,699	0	0	0	-0,699	0
analisis	0	0	0	0	1	1	-0,699	0	0	0	0	-0,699
banding	0	0	0	0	1	1	-0,699	0	0	0	0	-0,699
jalur	0	0	0	0	1	1	-0,699	0	0	0	0	-0,699
pendek	0	0	0	0	1	1	-0,699	0	0	0	0	-0,699
general	0	0	0	0	1	1	-0,699	0	0	0	0	-0,699
and	0	0	0	0	1	1	-0,699	0	0	0	0	-0,699
test	0	0	0	0	1	1	-0,699	0	0	0	0	-0,699
best	0	0	0	0	1	1	-0,699	0	0	0	0	-0,699
first	0	0	0	0	1	1	-0,699	0	0	0	0	-0,699
search	0	0	0	0	1	1	-0,699	0	0	0	0	-0,699

7) Topic Modeling* Menggunakan *Latent Dirichlet Allocation

Pembuatan *topic modeling* menggunakan *Latent Dirichlet Allocation*, menggunakan nilai TF-IDF yang telah diperoleh. *Teks* yang diprediksi topiknya dibatasi hanya menggunakan satu *teks* judul skripsi yang berasal dari T1. Berikut ini adalah nilai TF-IDF dari T1.

Tabel 10. Tabel Inisialisasi

Kata	Inisialisasi Kata	Penyebaran Kata (DF)
implementasi	1	4
metode	2	1
preference	3	1
ranking	4	1
organization	5	1
method	6	1
for	7	1
enrichment	8	1
evaluation	9	1
promethee	10	1
ii	11	1
tentu	12	3
naik	13	1
jabat	14	1
karyawan	15	1
pt	16	1
agung	17	1
sarana	18	1

Kemudian tentukan jumlah topik yang akan digunakan, dalam proses ini peneliti menggunakan 2 topik, yang di inisialisasi kan sebagai TA dan TB. Dilanjutkan dengan menentukan topik dari setiap kata yang terdapat pada *teks* judul skripsi secara acak. Berikut ini adalah pembagain kata berdasarkan topik nya.

Tabel 11. Tabel Pembagian Topik

TA		TB	
Kata	**DF**	**Kata**	**DF**
1	4	3	1
2	1	5	1
4	1	6	1
7	1	8	1
9	1	10	1
11	1	12	3
13	1	14	1
16	1	15	1
18	1	17	1

Berdasarkan penyebaran kata pada dua topik, maka dapat diketahui nilai probabilitas dari masing – masing topik, yaitu

Tabel 12. Tabel Probabilitas Topik

Teks	**Topik**			
	TA		**TB**	
	Frekuensi	**Probabilitas**	**Frekuensi**	**Probabilitas**
T1	12	**53%**	11	47%

Berdasarkan nilai probabilitas yang telah duperoleh dapat disimpulkan bahwa T1 memiliki topik TA dengan nilai probabilitas sebesar 53%.

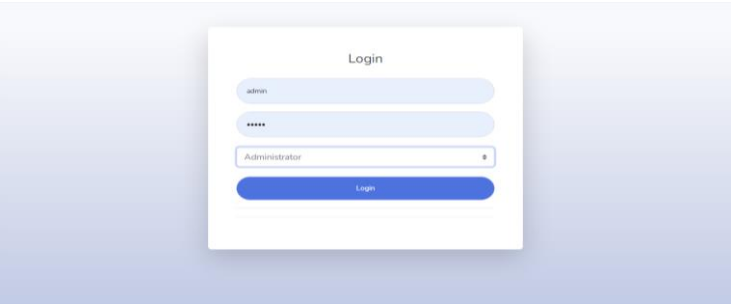
Berdasarkan proses analisis data serta representasi data, maka *topic modeling* menggunakan algoritma *Latent Dirichlet Allocation*, dapat membantu dalam menentukan topik terkait judul skripsi yang diajukan mahasiswa sehingga membantu proses penyeleksian judul skripsi pada program studi ilmu komputer.

c. Penerapan Sistem

Implementasi dari algoritma dan perancangan *user interface* yang telah disajikan pada sub bab sebelumnya, sebagai berikut :

1. Perancangan Halaman Login

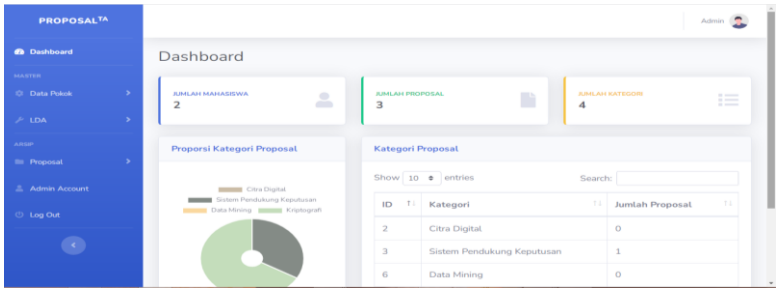
Halaman ini berisi *form login* dengan pilihan level login sebagai administrator dan mahasiswa. Berikut ini adalah desain tampilannya :



Gambar 2. Halaman Login

2. Perancangan Halaman Dashboard (Admin)

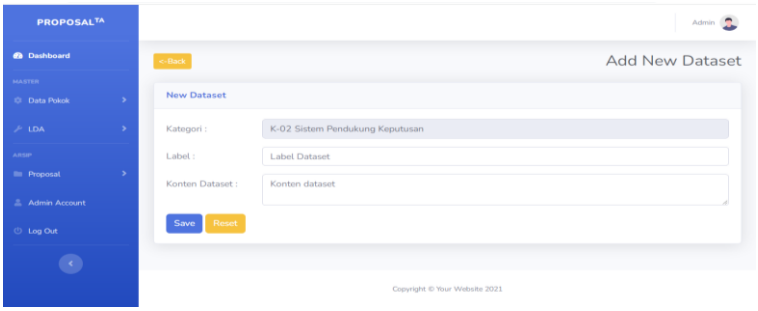
Halaman ini berisi informasi terkait jumlah mahasiswa, jumlah proposalan serta jumlah kategori yang telah masuk ke dalam *database* sistem.



Gambar 3. Halaman Dashboard (Admin)

3. Halaman *add new* dataset (Admin)

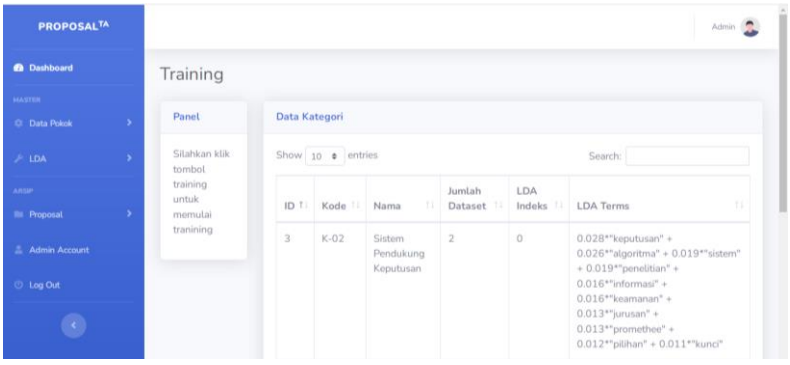
Halaman ini berfungsi untuk menambah *dataset* yang baru, sehingga jumlah kategori yang dapat dipilih lebih banyak variasinya.

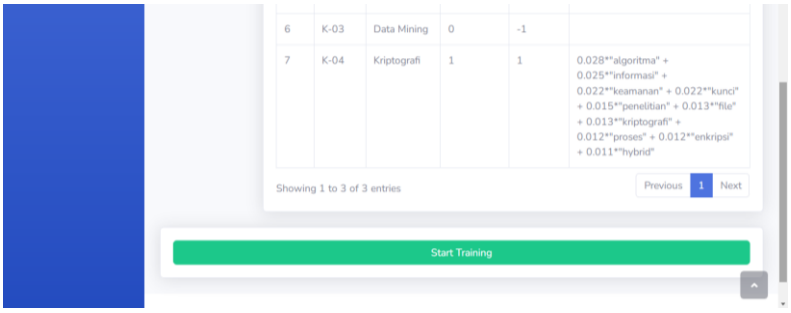


Gambar 4. Halaman New Dataset (Admin)

4. Halaman *Training* (Admin)

Pada halaman ini admin dapat melihat data kategori yang terdiri dari id, kode, nama, jumlah dataset, LDA indeks, dan LDA Term. Untuk melakukan *training* dapat dengan menekan tombol paling bawah yang tertera pada halaman ini.

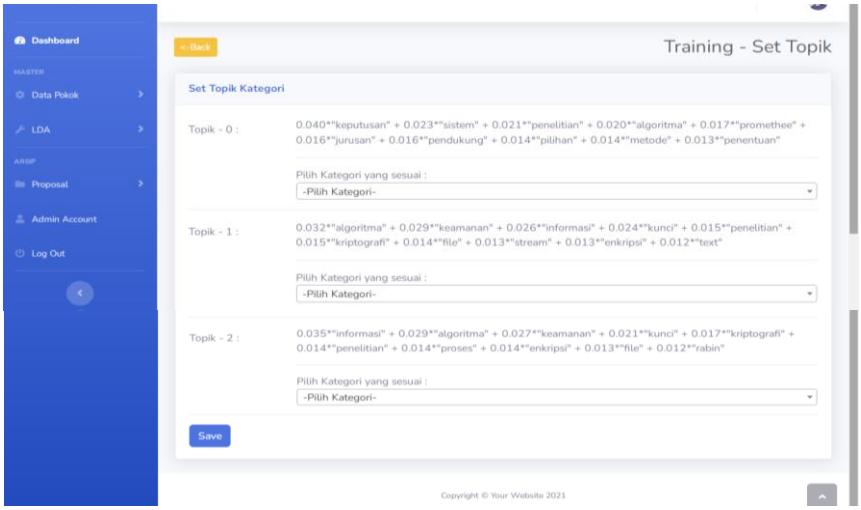




Gambar 5. Halaman Training Dataset (Admin)

5. Halaman *Set* Topik

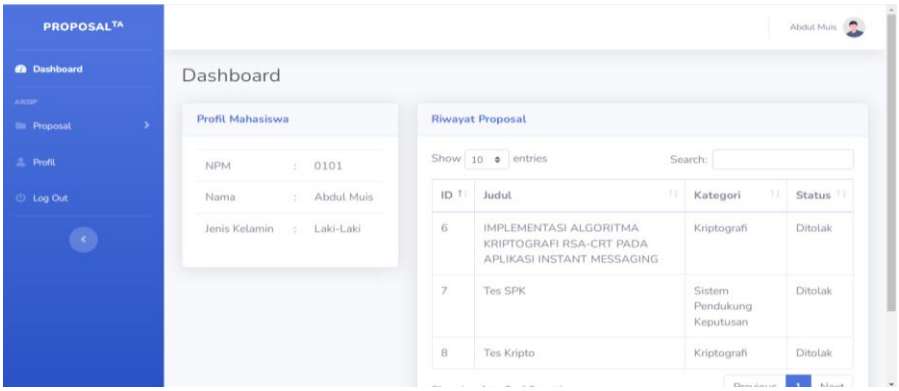
Pada halaman ini admin dapat menentukan topik dari setiap judul yang telah di daftarkan mahasiswa berdasarkan probabilitas dan LDA yang dimiliki oleh setiap judul



Gambar 6. Halaman Set Topik

6. Halaman Dashboard (Mahasiswa)

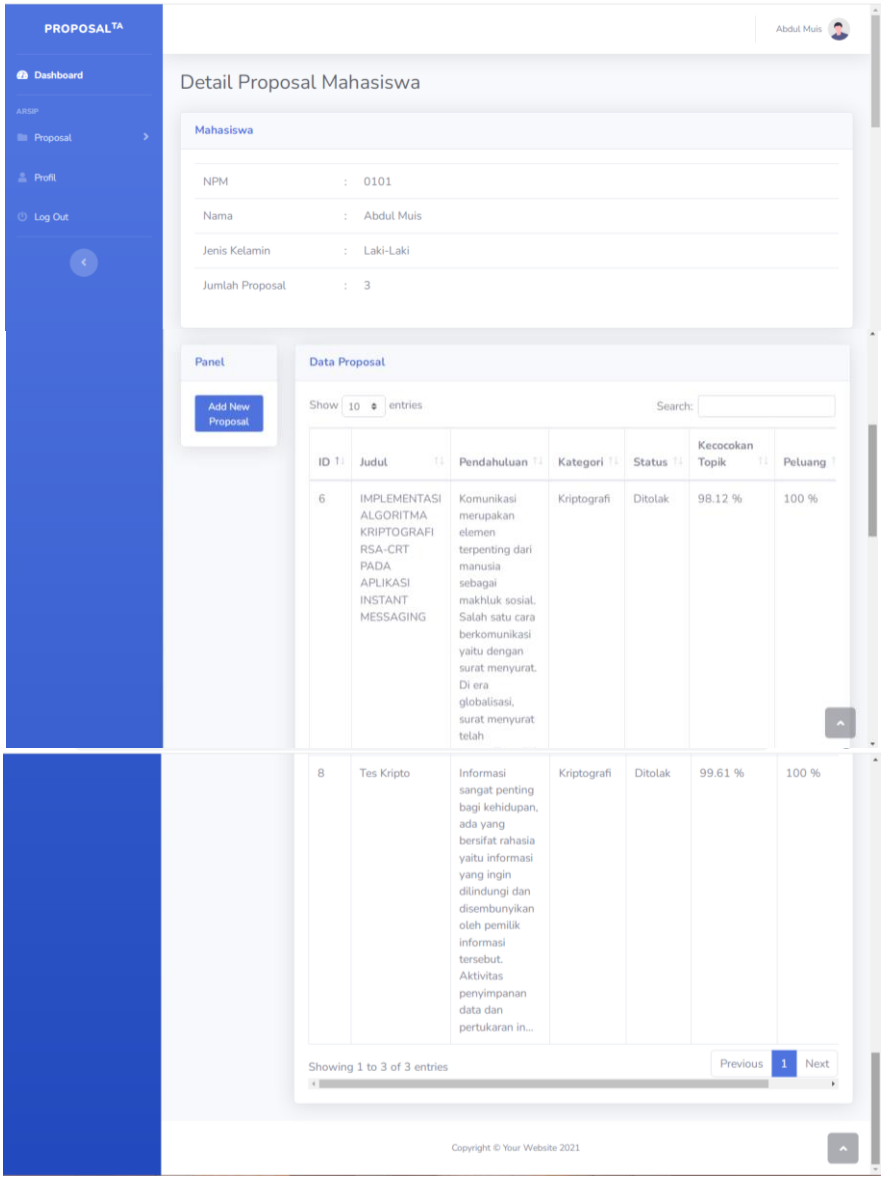
Pada halaman ini mahasiswa dapat meilihat data diri dan riwayat proposal yang telah diajukan.



Gambar 7. Halaman Dashboard (Mahasiswa)

7. Halaman Detail Proposal (Mahasiswa)

Pada halaman ini mahasiswa dapat melihat status dari judul skripsi yang telah diajukan secara detail, serta dapat mengetahui status dari judul yang diberikan dan melihat saran yang diberikan admin terkait judul yang diajukan apabila ada.



Gambar 8. Halaman Detail Proposal (Mahasiswa)

D. Simpulan

Berdasarkan hasil penelitian dengan judul "Penerapan Text Mining Pada Sistem Penyeleksian Judul Skripsi Mahasiswa Menggunakan Algoritma Latent Dirichlet Allocation Di Program Studi Ilmu Komputer UIN Sumatera Utara Medan" didapat kesimpulan bahwa algoritma *Latent Dirichlet Allocation* dapat digunakan untuk menentukan topik dari judul skripsi yang diajukan mahasiswa. Mahasiswa dapat mengetahui secara langsung kesesuaian topik dan peluang diterimanya setiap judul proposal yang diajukan. Dengan penggunaan sistem ini, proses

penyeleksian judul dapat di lakukan dalam waktu yang singkat.

E. Ucapan Terimakasih

Terimakasih kami sampaikan kepada LP2M Universitas Islam Negeri Sumatera Utara Medan yang telah menyelenggarakan kegiatan penelitian yang didanai oleh BOPTN sehingga terfasilitasinya kegiatan penelitian ini. Serta ucapan terimakasih kepada seluruh tim penelitian dan publikasi baik dari dosen, alumni dan mahasiswa yang membantu dalam mensukseskan dan menghasilkan penelitian ini.

F. Referensi

[1] Arikunto, S. (2017). *Prosedur Penelitian Suatu Pendekatan Praktek*.

[2] Balya. (2019). *Analisis Sentimen Pengguna Youtube di Indonesia pada Review Smartphone Menggunakan Naïve Bayes*. http://repositori.usu.ac.id/handle/123456789/23217

[3] Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

[4] Buntoro, G. A. (2017). Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. *Integer Journal*, *2*(1), 32–41. https://t.co/jrvaMsgBdH

[5] Gilang Kencana, C., & Sibaroni, Y. (n.d.). *Klasifikasi Sentiment Analysis pada Review Buku Novel Berbahasa Inggris dengan Menggunakan Metode Support Vector Machine (SVM)*.

[6] Gunawan, B., Pratiwi, H. S., & Pratama, E. E. (2018). Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)*, *4*(2), 113–118.

[7] Hakim, D. M. (2019). *Optical Music Recognition Pada Citra Notasi Musik Menggunakan Convolutional Neural Network*. UNIKOM.

[8] Kurniawan, D. (2021). *Pengenalan Machine Learning dengan Python* (Ed. 2). PT Elex Media Komputindo.

[9] Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2014). Foundations in Machine learning. In *SpringerBriefs in Computer Science* (Vol. 0, Issue 9783319056050).

[10] Muljono, Artanti, D. P., Syukur, A., Prihandono, A., & Setiadi, D. R. I. M. (2018). Analisa Sentimen Untuk Penilaian Pelayanan Situs Belanja Online Menggunakan Algoritma Naive Bayes. *Konferensi Nasional Sistem Informasi*, 165–170.

[11] Nugroho, D. D. A., & Alamsyah, A. (2018). Analisis Konten Pelanggan Airbnb Pada Network Sosial Media Twitter. *EProceedings of Management*, 1622–1628.

[12] Nurzahputra, A., & Muslim, A. (2016). Analisis Sentimen pada Opini Mahasiswa Menggunakan Natural Language Processing. *Seminar Nasional Ilmu Komputer*.

[13] Putra, K. B., & Kusumawardani, R. P. (2017). Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA). *Jurnal Teknik ITS*, *6*(2). https://doi.org/10.12962/j23373539.v6i2.23205

[14] Putu, I., Wirayasa, M., Made, I., Wirawan, A., Pradnyana, A., Kunci, K., Stemming, :, Bali, B., & Bastal, A. (n.d.). *ALGORITMA BASTAL: ADAPTASI ALGORITMA NAZIEF & ADRIANI UNTUK STEMMING TEKS BAHASA BALI* (Vol. 8).

[15] Rahim, R., Zufria, I., Kurniasih, N., Simargolang, M. Y., Hasibuan, A., Sutiksno, D. U., Nanuru, R. F., Anamofa, J. N., Ahmar, A. S., & Achmad Daengs, G. S. (2018). C4.5 classification data mining for inventory control. *International Journal of Engineering and Technology(UAE)*, *7*, 68–72. https://doi.org/10.14419/ijet.v7i2.3.12618

[16] Russell, S., & Norvig, P. (2021). Artificial Intelligence A Modern Approach (4th Edition). In *Pearson Series*.

[17] S, V., & R, J. (2016). Text Mining: open Source Tokenization Tools – An Analysis. *Advanced Computational Intelligence: An International Journal (ACII)*, *3*(1), 37–47. https://doi.org/10.5121/acii.2016.3104

[18] Sevsa, B. A., & Wahyudi, M. D. R. (2019). Analisis Sentimen pada Indeks Kinerja Dosen Fakultas SAINTEK UIN Sunan Kalijaga Menggunakan Naive Bayes Classifier. *Jurnal Buana Informatika*, *10*(2), 112–123.

[19] Sugiyono. (2016). Metode Penelitian Bisnis. Alfabeta: Bandung. *Jurnal Analisis*, *6*(2).

[20] Sunardi, Fadlil, A., & Suprianto. (2018). Analisis Sentimen Menggunakan Metode Naive Bayes Classifier Pada Angket Mahasiswa. *SAINTEKBU: Jurnal Sains Dan Teknologi*, *10*(2541–1942), 1–9.

[21] Ulfah Siregar, Z., Ruli, R., Siregar, A., & Arianto, R. (2019). *KLASIFIKASI SENTIMENT ANALYSIS PADA KOMENTAR PESERTA DIKLAT MENGGUNAKAN METODE K-NEAREST NEIGHBOR*. *8*(1).