PENERAPAN ALGORITMA *TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY* UNTUK REKOMENDASI PEMILIHAN JURUSAN MAHASISWA BARU

(Studi Kasus : Universitas Ichsan Gorontalo)

Oleh

PUTRI SITI SALSABILA IBRAHIM

T3119081

SKRIPSI

Untuk memenuhi salah satu syarat ujian

guna memperoleh gelar Sarjana

PROGRAM SARJANA

TEKNIK INFORMATIKA

UNIVERSITAS ICHSAN GORONTALO

2023

PERSETUJUAN USULAN PENELITIAN

PENERAPAN ALGORITMA *TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY* UNTUK REKOMENDASI PEMILIHAN JURUSAN MAHASISWA BARU

OLEH

PUTRI SITI SALSABILA IBRAHIM

T3119081

SKRIPSI

Untuk memenuhi salah satu syarat ujian
Guna memperoleh gelar sarjana
Program Studi Teknik Informatika
Ini telah disetujui oleh Tim Pembimbing
Gorontalo,

Pembimbing I	Pembimbing II
Hasdi, M.Kom	**Kartika Chandra Pelangi, M.Kom**
NIDN : 0907108701	NIDN : 0916038304

PENGESAHAN SKRIPSI

PENERAPAN ALGORITMA *TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY* UNTUK REKOMENDASI PEMILIHAN JURUSAN MAHASISWA BARU (Studi Kasus : Universitas Ichsan Gorontalo)

Oleh

PUTRI SITI SALSABILA IBRAHIM

T3119081

Diperkasa oleh Panitia Ujian Strata Satu (S1)

Universitas Ichsan Gorontalo

1. Ketua Penguji
   Irma Surya Kumala Idris, M.Kom .....
2. Anggota
   Sudirman Panna, M.Kom .....
3. Anggota
   Roys Pakaya, M.Kom .....
4. Anggota
   Husdi, M.Kom .....
5.. Anggota
   Kartika Chandra Pelangi, M.Kom .....

Mengetahui

Dekan Fakultas Ilmu Komputer	Ketua Program Studi
Irvan Abraham Salihi, M.Kom	**Sudirman S Panna, M.Kom**
NIDN. 0928028101	NIDN. 0924038205

PERNYATAAN SKRIPSI

Dengan ini saya menyatakan bahwa :

1. Karya tulis (Skripsi) saya ini adalah asli dan belum pernah diajukan untuk mendapatkan gelar akademik (Sarjana) baik di Universitas Ichsan Gorontalo maupun diperguruan tinggi lainnya.
2. Karya tulis (Skripsi) saya ini adalah murni gagasan, rumusan, dan penelitian saya sendiri, tanpa bantuan pihak lain, kecuali arahan dai Tim Pembimbing.
3. Dalam karya tulis (Skripsi) saya ini tidak terdapat karya atau pendapat yang telah dipublikasikan orang lain, kecuali secara tertulis dicantum sebagai acuan/sitasi dalam naskah dan dicantumkan pula dalam daftar pustaka.
4. Peryataan ini saya buat dengan sesungguhnya dan apabila dikemudian hari terdapat penyimpangan dan ketidakbenaran dalam peryataan ini, saya bersedia menerima sanksi akademi berupa pencabutan gelar yang telah diperoleh karena karya tulis ini, serta sanksi lainnya sesuai dengan norma- norma yang berlaku di Universitas Ichsan Gorontalo.

Kota, Mei 2023
Yang Membuat Peryataan,

Putri Siti Salsabila Ibrahim

ABSTRACT

PUTRI SITI SALSABILA IBRAHIM. T3119081. APPLICATION OF TERM FREQUENCY - INVERSE DOCUMENT FREQUENCY ALGORITHM FOR RECOMMENDATION OF NEW STUDENT DEPARTMENT SELECTION (A CASE STUDY OF UNIVERSITAS ICHSAN GORONTALO)

Admission of new students or commonly referred to as PMB is a routine activity held by universities every opening of the new academic year. Universitas Ichsan Gorontalo until now still has difficulty in recognizing whether one student is capable or not in the department he has chosen. With an interview session conducted by the lecturers, people will find which department is most suitable for the student. This research focuses on the main application of the calculation of TF value and IDF value of each keyword to each document to be processed and is expected to provide accurate results to find out which department is most suitable for the new students. After the classification process using the TF.IDF and K-Nearest Neighbor algorithms with several trials of the K value, the K = 6 and K = 8 values are obtained with the evaluation results using the Confusion matrix method, calculating a good accuracy rate of 70%.

Keywords: new student, TF.IDF, KNN, Confusion Matrix

ABSTRAK

PUTRI SITI SALSABILA IBRAHIM. T3119081. PENERAPAN ALGORITMA *TERM FREQUENCY - INVERSE DOCUMENT FREQUENCY* UNTUK REKOMENDASI PEMILIHAN JURUSAN MAHASISWA BARU (STUDI KASUS UNIVERSITAS ICHSAN GORONTALO)

Penerimaan mahasiswa baru atau biasa disebut dengan PMB yang merupakan aktivitas rutin diadakan oleh perguruan tinggi setiap pembukaan ajaran baru. Universitas Ichsan Gorontalo hingga saat ini masih kesulitan dalam mengetahui apakah mahasiswa ini mampu tidak pada jurusan yang telah dipilihnya. Dengan adanya sesi wawancara yang dilakukan oleh dosen- dosen pengajar dapat kita ketahui jurusan mana yang paling sesuai dengan mahasiswa tersebut. Penelitian ini berfokus pada penerapan utama dari perhitungan nilai TF dan nilai IDF dari setiap kata kunci terhadap masing – masing dokumen yang akan di olah, dan diharapkan dapat memberikan hasil yang akurat untuk dapat mengetahui jurusan mana yang paling sesuai dengan mahasiswa baru tersebut. Setelah dilakukan proses klasifikasi menggunakan algoritma TF.IDF dan K-Nearest Neighbor dengan beberapa kali uji coba nilai K, didapatkan Nilai K=6 dan K=8 dengan hasil evaluasi menggunakan metode Confusion matrix perhitungan tingkat akurasi yang baik sebesar 70%.

Kata kunci: mahasiswa baru, TF.IDF, KNN, *Confusion Matrix*

KATA PENGANTAR

Alhamdulillah, penulis dapat menyelesaikan hasil penelitian ini dengan judul "Penerapan Algoritma *Term Frequency – Inverse Document Frequency* untuk rekomendasi pemilihan jurusan mahasiswa baru" untuk memenuhi salah satu syarat penyusunan Skripsi Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Ichsan Gorontalo.

Penulis menyadari sepenuhnya bahwa usulan penelitian ini tidak mungkin terwujud tanpa bantuan dan dorongan dari berbagai pihak, baik bantuan moril maupun materil. Untuk itu, dengan segala keikhlasan dan kerendahan hati, penulis mengucapkan banyak terima kasih dan penghargaan yang setinggi-tingginya kepada :

1. Muh. Ichsan Gaffar, SE.,M.Ak, selaku ketua Yayasan Pengembangan Ilmu Pengetahuan dan Teknologi (TPIPT) Ichsan Gorontalo
2. Dr. Abdul Gaffar La Tjokke, M.Si, Selaku Rektor Universitas Ichsan Gorontalo
3. Irvan A. Salihi, M.Kom, Selaku Dekan Fkultas Ilmu Komputer Universitas Ichsan Gorontalo
4. Sudirman Melangi, M.Kom, Selaku pembantu Dekan I Bidang Akademik Fakultas Ilmu Komputer Universitas Ichsan Gorontalo
5. Irma Surya Kumala, M.Kom selaku Pembantu Dekan II Bidang Administrasi Umum dan Keuangan Fakultas Ilmu Komputer Universitas Ichsan Gorontalo
6. Sudirman S. Panna, M.Kom, Selaku ketua jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Ichsan Gorontalo
7. Husdi, M.Kom , Selaku Pembimbing I
8. Kartika Chandra Pelangi, S.Kom., M.Kom, Selaku Pembimbing II
9. Bapak dan Ibu Dosen Universitas Ichsan Gorontalo yang telah mendidik dan mengajarkan berbagai disiplin ilmu kepada penulis
10. Kedua Orang Tua saya yang tercinta, atas segala kasih sayang, jerih payah dan doa restunya dalam membesarkan dan mendidik penulis
11. Rekan-rekan seperjuangan yang telah banyak memberikan bantuan dan mendukung moril yang sangat besar kepada penulis
12. Kepada semua pihak yang ikut membantu dalam penyelesaian proposal/skripsi ini yang tidak sempat penulis sebut satu-persatu.

Semoga Allah SWT melimpahkan balasan atas jasa-jasa mereka kepada kami.

Penulis menyadari sepenuhnya bahwa apa yang telah dicapai ini masih jauh dari kesempurnaan dan masih banyak terdapat kekurangan. Oleh karena itu, penulis sangat mengharapkan adanya kritik dan saran yang konstruktif. Akhirnya penulis berharap semoga hasil yang telah dicapai ini dapat bermanfaat bagi kita semua, Amiin.

Gorontalo, Mei 2023

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

BAB I

PENDAHULUAN

1.1 Latar Belakang

Universitas Ichsan merupakan salah satu perguruan tinggi swsta yang berada di provinsi gorontalo yang berdiri sejak tahun 1988 di kenal dengan kursus Akuntansi hingga saat ini sudah berkembang menjadi suatu Universitas, Unisan merupakan Universitas yang terdaftar dan teregistrasi di gorontalo sebagai Universitas Terbaik, Tidak hanya Institusi yang terakreditasi baik tetapi beberapa program studinyapun sudah terakreditas Baik.Perguruan tinggi ini berskala nasional dan masuk 10 besar perguruan tinggi swasta se LLDIKTI wilayah IX Sulawesi, dan juga memiliki tenaga pendidik dengaan standar kualitas internasional dan professional, dengan di dukung sarana dan prasarana yang tersedia, sehingga dapat menunjang kegiatan pembelajaran antar mahasiswa dan dosen pengajar.

Penerimaan mahasiswa baru atau biasa disebut dengan PMB yang merupakan aktivitas rutin diadakan oleh perguruan tinggi setiap pembukaan ajaran baru. Dengan system pendaftaran secara Online maupun Offline. Pada Proses Pendaftaran ini, calon mahasiswa diharuskan mengisi Biodata dan juga menyediakan beberapa persyaratan yang perlu di siapkan. Selanjutnya petugas Administrasi hanya mengolah layout folmulir, kemudian mengecek data pendaftaran mahasiswa yang mendaftar kemudian di input pada computer agar calon mahasiswa tersebut dapat memiliki nomor pendaftaran dan juga resmi terdaftar sebagai calon mahasiswa baru pada Universitas Ichsan. Berikut beberapa program studi yang terdapat pada universitas ichsan, serta kuota tiap prodi yang disediakan, dapat dilihat pada Tabel 1.1

Tabel 1.1 Kuota Program Studi T.A 2023/2024

Fakultas	Program Studi	Jumlah
Ekonomi	Manajemen	180

	Akuntansi	120
Hukum	Ilmu Hukum	120
Ilmu Komputer	Teknik Informatika	240
	Sistem Informasi	120
	DKV	60
Pertanian	Agroteknologi	60
	Agribisnis	60
	THP	60
Teknik	Teknik Elektro	60
	Teknik Arsitektur	90
Fisip	Ilmu Komunikasi	80
	Ilmu Pemerintahan	120

Sumber: Panitia penerimaan mahasiswa baru 2023/2024

Universitas ichsan hingga saat ini masih kesulitan dalam mengetahui apakah mahasiswa ini mampu tidak pada jurusan yang telah di pilihnya, Karena terkait pemilihan program studi ini terdapat 87% mahasiswa Indonesia salah dalam memilih jurusan.[Irene Guntur], Maka dari itu unisan mengutamakan penilaiannya dalam menerima mahasiswa baru dilihat dari minat jurusan pilihan mahasiswa, dengan adanya sesi wawancara yang dilakukan oleh dosen – dosen pengajar dapat kita ketahui jurusan mana yang paling sesuai dengan mahasiswa tersebut. Banyaknya mahasiswa yang melakukan tes sehingga proses seleksi yang dikerjakan oleh panitia dalam melakukan perbandingan mahasiswa tersebut layak atau tidak dalam melanjutkan pada program studi yang dipilihnya cukup memakan waktu.

Berdasarkan interview yang dilakukan kepada wakil ketua panitia maba sekaligus dosen yang ikut serta dalam melakukan wawancara terhadap calon mahasiswa baru, dikatakan bahwa benar masih banyak mahasiswa yang bingung dalam memilih jurusan baik itu pada saat pendaftaran maupun pada saat tes wawancara di adakan, disebutkan beberapa faktor calon mahasiswa baru bingung dalam memilih jurusan yaitu, karena perbedaan jurusan pada saat di sekolah dengan jurusan yang akan di pilihnya, selnjutnya karena hanya ikut ikutan temannya, terakhir yaitu termasuk juga karena impian dan kemauan orang tua.

Beberapa penelitian yang telah menggunakan berbagai macam pendekatan dengan berbagai algoritma yang berbeda untuk memberikan solusi dalam mengetahui jurusan mana yang paling sesuai dengan mahasiswa tersebut, contoh algoritmanya seperti, C4.5, K-Means,TF-IDF, dan Naïve Bayes. metode yang di pergunakan pada kasus ini adalah TF-IDF, dengan tahapan processing text mining yang dilakukan mempersiapan dokumen dan seleksi dokumen karena Tanpa Pocessing text mining pembobotan kata tidak dapat dilakukan. TF-IDF juga di gunakan karena paling baik dalam memperoleh informasi dan untuk mendapatkan hasil yang optimal maka di gunakanlah Algoritma ini [1]

Metode *Term Frequency – Inverse document Frequency* (TF-IDF) Merupakan algoritma yang digunakan untuk menganalisa hubungan bobot suatu term terhadap dokumen [2] seperti pada penelitian sebelumnya yang menggunakan metode TF-IDF mengenai "Penerapan Algoritma *Cosine Similarity* dan Pembobotan TF-IDF System Penerimaan Mahasiswa Baru pada Kampus Swasta", "Implementasi Algoritma TF-IDF untuk pencarian pedoman Akademik dan penentuan sanksi pada jurusan Teknik Informatika UIN sunan Gunung djati Bandung", dan "Penerapan Algoritma *Term Frequency – Inverse Dodument Frequency* Untuk text Mining". Beberapa penelitian sebelumnya ini menggunakan pembobotan TF-IDF karena dapat memberikan nilai bobot pada setiap kata dalam dokumen. Maka dari itu metode ini tepat untuk mengelolah data tes yang di isi oleh mahasiswa baru.

Pada proses perhitungan dari algoritma TF.IDF untuk penerapan algoritma dalam menghitung bobot pada suatu data digunakan suatu tahapann *preprocessing* untuk menghasilkan data yang secara teratur. Dalam *preprocessing* atau (*Natural Language Processing)* adalah suatu proses dimana melakukan pelabelan secara otomatis dengan suatu Teknik berupa *case folding,tokenizing,stopword, dan steaming* dengan menggunakan *tools* pada python dengan *Library streamlit* untuk memproses tahapan *preprocessing* dengan menggunakan *Natural Language Tools Kit* (NLTK). Pada penggunaan python untuk tahapan preprocessing NLTK digunakan untuk mengolah data dalam jumlah yang banyak untuk menganalisis data.

Berdasarkan Uraian diatas, peneliti merasa tertarik untuk menerapkan metode *Term Frequency – Inverse Document Frequency* (TF-IDF) dan akan mencoba untuk mempraktekan metode algoritma ini dalam penelitian dengan judul **"Penerapan Algoritma *Term Frequency – Inverse Document Frequency* Untuk Rekomendasi Pemilihan Jurusan Mahasiswa Baru"** Penelitian ini berfokus pada penerapan utama dari perhitungan nilai TF dan nilai IDF dari setiap kata kunci terhadap masing – masing dokumen yang akan di olah,

dan diharapkan dapat memberikan hasil yang akurat untuk dapat mengetahui jurusan mana yang paling sesuai dengan mahasiswa baru tersebut dan dapat menjadi pertimbangan untuk tugas akhir.

1.2 Identifikasi Masalah

Proses mengklasifikasi yang dilakukan secara manual akan memakan waktu dan tenaga, sehingga diperlukan suatu metode untuk mempermudah dan membantu proses pengklasifikasi hasil dari pernyataan calon mahasiswa terhadap data teks hasil wawancara,sehingga dapat memperoleh mahasiswa tersebut cocok pada jurusan apa

1.3 Rumusan Masalah

1. Bagaimana menerapkan Algoritma *Term Frequency – Inverse Document Frequency* (TF-IDF) untuk rekomendasi pemilihan jurusan pada mahasiswa baru?
2. Bagaimana hasil penerapan Algoritma *Term Frequency – Inverse Document Frequency* (TF-IDF) untuk rekomendasi pemilihan jurusan pada mahasiswa baru?

1.4 Batasan Masalah

Data yang digunakan berupa data teks dari tes wawancara Mahasiswa Universitas Ichsan Gorontalo.

1.5 Tujuan Penelitian

1. untuk mengetahui cara kerja Algoritma Term Frequency – Inverse Document Frequency (TF-IDF)
2. Untuk mengetahui hasil penerapan *Algoritma Term Frequency – Inverse Document Frequency* (TF-IDF)

1.6 Manfaat Penelitian

1. Manfaat Teoritis: Memberikan masukan perkembangan bagi pengembangan ilmu pengetahuan dan teknologi , khususnya pada bidang penerimaan mahasiswa baru, yaitu berupa uji coba metode Term Frequency – Inverse Document Frequency.
2. Manfaat Praktis: Sehubungan pemikiran, karya bahan pertimbangan, atau solusi yang dapat dijadikan salah satu dasar yang tepat dan akurat dalam melakukan penelitian penerapan algoritma TF-IDF untuk mengetahui jurusan mana yang paling sesuai dan layak untuk mahasiswa baru tersebut, dan agar dapat memberikan kontribusi untuk bidang penerimaan mahasiswa baru.

BAB II
LANDASAN TEORI

2.1 Tinjauan Studi

Yang menjadi tinjauan studi dalam penelitian ini adalah sebagai berikut:

Tabel 2.1 Tinjauan Studi

No	Peneliti	Judul	Tahun	Metode	Hasil
1	Apriani, Hizbu zakiyudin,Khairan Marzuki [3] .	Penerapan Algoritma *Cosine Similarity* dan Pembobotan TF-IDF System Penerimaan Mahasiswa Baru pada Kampus Swasta	2021	*Cosine Similarity* TF-IDF	Dengan menggunakan kombinasi pembobotan TF-IDF dan metode *Cosine Similarity* bisa mendapatkan tingkat akurasi hingga 64,28.
2	Nico,Utomo Budiyanto, Titin Fatimah [4].	Implementasi Algoritma Pembobotan TF-IDF dan *Cosine Similarity* untuk Penetapan Kategori Artikel pada	2022	*Cosine Similarity* TF-IDF	Penelitian ini berhasil membuat alat bantu berupa perangkat lunak yang bertujuan untuk mengklasifikasi kategori artikel di website

		website Universitas Budi Luhur			Universitas Budi Luhur.
3	Musfiroh Nurjannah, Hamdani,Inda fitri astute [2].	Penerapan Algoritma *Term Frequency – Inverse Document Frecuency* untuk *Text Mining*	2021	*Term Frequency – Inverse Document Frecuency*	Berdasarkan hasil peneliti dapat menyimpulkan bahwa dengan menggunakan kombinasi pembobotan TF-IDF dan metode *Cosine Similarity* bisa mendapatkan tingkat akurasi hingga 64,28.

2.2 Tinjauan Pustaka

2.2.1 Mahasiswa Baru

Pada tahun pertama perkuliahan mahasiswa menyandang status sebagai mahasiswa baru, dalam hal ini menjadi seorang mahasiswa baru memiliki peran dan tantangan sendiri, karena sudah di anggap bisa bertanggung jawab atas dirinya sendiri dalam menghadapi berbagai tantangan, diharapkan mahasiswa dapat menghadapi lika liku problem dalam menjadi mahasiswa karena keberhasilan seorang mahasiswa baru dalam bangku perkuliahan terdapat pada proses belajarnya [5].

2.2.2 Jurusan

Jurusan ialah suatu bagian dari fakultas, yaitu suatu pelajaran yang di pilih atau di tekuni oleh mahasiswa, dalam satu fakultas terdapat beberapa jurusan yang berbeda pelajarannya, contohnya seperti pada fakultas ilmu komputer Unisan terdapat jurusan Teknik Informatika, Seistem Informasi, dan Desain Komunikasi Visual.

2.2.3 Python

Python adalah Bahasa pemrograman yang mempunyai banyak kegunaan karena berfokus pada tingkat dari pembacaan kode sehingga saat ini Python menjadi salah satu bahasa pemrograman yang menjadi pilihan dari banyak programmer dalam menyelesaikan pekerjaannya. Python mempunyai sintaksis kode yang sangat jelas dan juga dilengkapi fungsionalitas yang tinggi sehingga mempermudah dalam meyelesaikan pekerjaan terkait pemrograman. Python merupakan bahasa pemrograman yang dirancang dengan memeprhatikan kemudahan programmer agar bisa dengan lebih efisien mengerjakan pekerjaan. Meskipun terkenal dengan bahasa pemrograman yang mudah dan efisen akan tetapi bahasa pemrograman ini termasuk pada high level

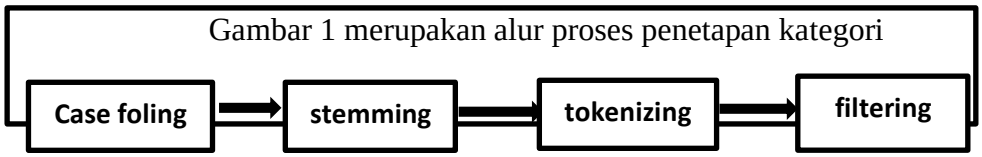
2.2.4 Analisis Data

Data yang digunakan pada penelitian ini berupa data text yang berisikan tentang penerimaan mahasiswa baru di Universitas Ichsan Gorontalo, terkait dengan jurusan apa yang akan di ambil, mengapa memilih jurusan ini, dan pertanyaan lainnya. metode pengumpulan data yaitu dengan wawancara yang dilakukan oleh Panitia Penerimaan Mahasiswa Baru (PMB) dan juga dosen-dosen pengajar lainnya

2.2.5 Text Mining

Text Mining adalah sebutan yang mendeskripsikan kumpulan data teks semi-terstruktur, maupun tidak terstruktur yang bertujuan untuk mendapatkan informasi yang berguna dari beberapa sekumpulan dokumen [1], [4]. Dengan adanya bantuan dari *Text Mining* ini, sehingga suatu dokumen dapat diketahui jenis kategorinya melalui kata yang terdapat pada dokumen tersebut [6].

Dengan demikian *Text Mining* didefinisikan sebagai suatu proses untuk mendapatkan informasi dimana pengguna dapat berinteraksi dengan sekumpulan dokumen dengan menggunakan beberapa tahapan yang disebut dengan *Text Processing* [7]. Menggunakan *tools* yang ada pada pyton dengan Library (*Natural Language Tool Kit)* NLTK ialah suatu platform yang banyak digunakan untuk membangun analisis teks.



Gambar 2.1 Proses *Ekstrasi* Dokumen

a. *Case Foling*

Case Folding merupakan bentuk *text* yang efektif dan paling sederhana, tujuannya untuk mengubah huruf besar yang berada dalam dokumen menjadi huruf kecil. dan hanya akan menerima huruf "a" sampai huruf "z" selain karakter huruf abjad akan dihilangkan [1].

b. Tokenizing

Merupakan suatu proses yang dilakukan untuk memisahkan dokumen dari suatu kalimat menjadi kumpulan bagian – bagian kata tertentu, pada proses ini juga menghilangkan karakter tertentu seperti tanda baca dan lainnya [2].

c. Filtering

Pada tahap ini menjelaskan proses mengubah atau menyaring kata menggunakan algoritmma *Stopword* Yaitu dengan menghilangkan suatu kata yang tidak mempunyai makna penting untuk dijadikan sebagai suatu kata kunci [8]. Seperti kata “atau”, “di”, “dan”, “yang” dan beberapa kata lainnya.

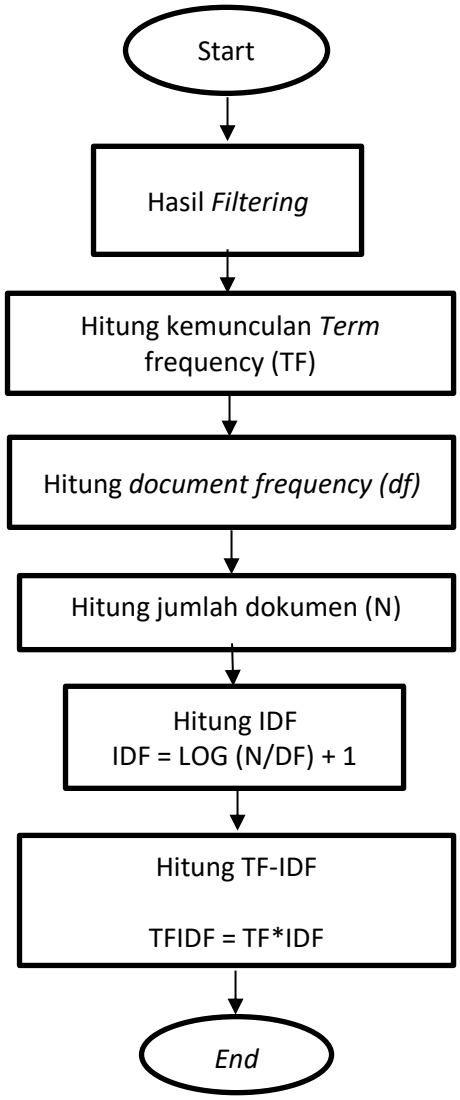
d. Stemming

Menjelaskan tentang bagaimana menghilangkan atau mengubah suatu kata menjadi kata dasar, atau disebut dengan pembentukan kata baru tetapi tidak mengubah kelas katanya, seperti “berjalan” menjadi “jalan” [4].

2.2.5 Algoritma *Term Frequency – Inverse Document Frequency* (TF-IDF)

Term Frequency – Inverse Document Frequency Merupakan suatu metode yang digunakan untuk menghitung Suatu nilai bobot kata (*Term)* terhadap dokumen yang akan di olah [3]. Metode ini digunakan dengan menggabungkan dua konsep untuk menghitung bobot, yaitu nilai frekuensi kemunculan sebuah kata yang disebut dengan *Term Frequency* (TF) dalam suatu dokumen tertentu dan *Inverse frekuensi* dokumen yang terdapat kata yang disebut dengan *Inverse document frequency* (IDF) [1].

Berikut adalah gambar tahapan *Term Frequency – Inverse Document Frequency*



Gambar 2.2 Tahapan pembobotan TFIDF

a. ***Term Frequency* (TF)**

TF atau yang disebut dengan *Term Frequency* adalah bobot dari suatu kata "t" dalam dokumen "d" yang dilambangkan dengan tf.idf. Maksud dari Tf yaitu jika semakin tinggi nilai frekuensi kemunculan kata *(Term)* pada suatu dokumen maka akan semakin tinggi juga bobot nilai untuk *Term* itu sendiri [3].

b. ***Document Frequency* (DF)**

Documen Frequency adalah jumlah suatu dokumen yang mengandung kata *(term),* Pada setiap kata akan dihitung nilai *Document Frequency* (DF), selain itu (DF) juga merupakan metode *Feature selection* yang paling sederhana dengan waktu komputasi yang bisa di bilang rendah [4].

c. ***Inverse Document Frequency* (IDF)**

Inverse Document Frequency merupakan kebalikan dari suatu suatu proses (TF), Maksudnya yaitu jika semakin tinggi /frekuensi kemunculan kata *(term)* maka nilai dari bobot kata itu sendiri akan semakin kecil [4].

Ada berapapun besaran nilai tfid, makan apabila N = n, akan d idapatkan hasil (nol), dalam hal ini dikarenakan hasil dari log 1. Sehingga untuk menghitung IDF ditambahkan nilai 1 (satu) pada samping IDF, sehingga cara kerja perhitungan bobotnya seperti pada rumus (1).

$$W_{ij} = tf \times idf$$

$$W_{ij} = tf_{ij} \times \left(\log \frac{N}{n} + 1\right) \quad (1)$$

Gambar 2.3 Rumus TF-IDF

Dengan keterangan sebagai berikut:

Wij	= bobot *term*/kata tj terhadap dokumen di
Tfij	= jumlah muncul *term*/kata t pada dalam dokumen d
N	= Total keseluruhan dokumen
n	= Jumlah dokumen yang mengandung *term*/kata tj

2.2.6 Penerapan Algoritma TF-IDF

Dokumen yang digunakan dalam uji coba ini menggunakan dokumen Tes wawancara mahasiswa baru yaitu berupa Teks,

Contoh *Implementasi* dari algoritma *Term frequency – Inverse document frequency* yaitu menggunakan dokumen yang telah dilakukan teks *processing* dibawah sebagai *Query*

Tahap awal yaitu dilakukan *preprocessing* dari semua dokumen D1, D2, D3. *Preprocessing* dilakukan menggunakan *case foling, steaming, tokenizing,* dan *filtering.* Pada table dibawah merupakan table hasil dari *preprocessing* dokumen Q [3].

Tabel 2.2 *Ekstrasi* dokumen 1 (D1)

Case foling	**steaming**	**tokenizing**	**Filtering**
saya sangat tertarik masuk jurusan hukum, karena saya ingin menjadi seorang ahli hukum, dan ingin jadi pengacara	saya sangat tertarik masuk jurusan hukum karena saya ingin jadi seorang ahli hukum, dan ingin jadi pengacara	‘saya’,’sangat’,’tertarik’,’masuk’,’jurusan’,’hukum’,’karena’,’saya’,’ingin’,’jadi’,’seorang’,’ahli’,’hukum’,’dan’,’ingin’,’jadi’,’pengacara’	’saya’,’sangat’,’tertarik’,’masuk’,’jurusan’,’hukum’,’ingin’,’jadi’,’orang’,’ahli’,’pengacara’,’orang’,’pengacara’,’publik’,’speaking’,’bisa’,’yakin’,’orang’,’ilmu’,’pengetahuan’,’sosial’
ingin menjadi seorang pengacara	ingin jadi orang pengacara	ingin’,’jadi’,’orang’,’pengacara’	
publik speaking dan bisa meyakinkan orang	publicspeaking dan bisa yakin orang	‘publik’, ’speaking’,’bisa’,’yakin’,’orang	
ilmu pengetahuan social	ilmu pengetahuan social	ilmu pengetahuan sosial	

Tabel 2.3 *ekstrasi* dokumen 2 (D2)

Case foling	**steaming**	**tokenizing**	**filtering**
menurut saya kampus ini cukup bagus, di bidang it oleh karena itu saya mengambil jurusan informatika, karena saya tertarik di bidang it dan computer	menurut saya kampus ini cukup bagus di bidang it oleh karena itu saya ambil jurusan teknik informatika, karena saya tertarik di bidang it dan komputer	‘menurut’,’say a’,’kampus’,’i ni’,’cukup’,’ba gus’,’di’,’bida ng’,’it’,’oleh’,’ karena’,’itu’,’s aya’,’ambil’,’j urusan’,’teknik ’,’informatika’ ,’karena’,’saya ’,’tertarik’,’di’, ’bidang’,’it’,’d an’,’komputer’	‘menurut’,’say’,’ kampus’,’cukup’ ,’bagus’,’bidang’ ,’it’,’karena’,’say a’,’ambil’,’jurus an’,’teknik’,’info rmatika’, bidang’,’komput er’cita-cita’,’ingin’,’jadi ’,’bisnis’punya’,’ lebih’,’hapal’tek nik’,’komputer’,’ jaringan’
cita-cita saya ingin menjadi pembisnis di bidang it	cita-cita saya ingin jadi bisnis di bidang it	‘cita-cita’, ’saya’,’ingin’,’ jadi’,’bisnis’,’ di’,’bidang’,’it ’,’	
saya punya kelebihan menghapal	saya punya lebih hapal	‘saya’,’punya’, ’lebih’,’hapal	
teknik komputer jaringan	teknik komputer jaringan	‘teknik’,’komp uter’,’jaringan’	

Tabel 2.4 *ekstrasi* dokumen 3 (D3)

Case foling	**steaming**	**tokenizing**	**filtering**
ingin mengasah kemampuan pengetahuan lebih mendalam, ingin	ingin asah mampu pengetahuan	‘ingin’,’asah’,’ mampu’,’peng etahuan’,’lebih	‘ingin’,’asah’,’ma mpu’,’pengetahua n’,’lebih’,’dalam’

menjadi salah satu pembisnis ternama di Indonesia	lebih dalam ingin jadi salah satu bisnis nama di Indonesia	’,’dalam’,’ingin’,’jadi’,’salah’,’satu’,’bisnis’,’nama’,’indonesia’	,’jadi’,’salah’,’satu’,’bisnis’,’nama’,’indonesia’,’pengusaha’,’sendiri’,’belum’,’lebih’,’saya’,’bisa’,’dalam’,’hitung’,’akuntansi’,
ingin jadi pengusaha	ingin jadi pengusaha	‘ingin’,’jadi’,’pengusaha	
saya sendiri belum tau apa kelebihan saya, tetapi saya bisa dalam menghitung	saya sendiri belum tau apa kelebihan saya tetapi saya bisa dalam hitung	‘saya’,’sendiri’,’belum’,’tau’,’apa’,’kelebihan’,’saya’.’tetapi’,’saya’,’bisa’,’dalam’,’hitung’	
Akuntansi	Akuntansi	‘akuntansi’	

Setelah *preprocessing* dilakukan, proses yang akan dilakukan selanjutnya yaitu, melakukan perhitungan untuk mengetahui bobot perkata engan menghitung jumlah *term frequency* dokumen (tf).

Langkah 1 Menghitung Term Frequency(tf)

Tabel 2.5 Menghitung TF

Menghitung Term Frequency(tf)

No	Term	D1	D2	D3
1	Saya	1	1	1
2	Sangat	1	0	0
3	Tertarik	1	1	0
4	Masuk	1	0	0
5	Jurusan	1	0	0
6	Hokum	1	0	0
7	Karena	1	1	0
8	Ingin	1	1	1
9	Jadi	1	1	1
10	Orang	1	0	0
11	Ahli	1	0	0
12	Hokum	1	0	0
13	Pengacara	1	0	0
14	Public	1	0	0
15	Speaking	1	0	0
16	Bias	1	0	0
17	Yakin	1	0	0
18	Ilmu	1	0	0
19	Pengetahuan	1	0	1
20	Social	1	0	0
21	Menurut	0	1	0
22	Cukup	0	1	0
23	bagus	0	1	0
24	Bidang	0	1	0
25	It	0	1	0
26	ambil	0	1	0
27	informatika	0	1	0
28	komputer	0	1	0
29	cita-cita	0	1	0
30	bisnis	0	1	1
31	punya	0	1	0
32	lebih	0	1	1
33	hapal	0	1	0
34	teknik	0	1	0
35	jaringan	0	1	0
36	asah	0	0	1
37	kemampuan	0	0	1
38	dalam	0	0	1
39	salah	0	0	1
40	satu	0	0	1
41	ternama	0	0	1
42	indonesia	0	0	1
43	pengusaha	0	0	1
44	sendiri	0	0	1
45	belum	0	0	1
46	tau	0	0	1
47	apa	0	0	1
48	tetapi	0	0	1
49	bisa	0	0	1
50	hitung	0	0	1
51	akuntansi	0	0	1

Langkah ke 2 Menghitung Document Frequency maksudnya yaitu melakukan perhitungan nilai jumlah dokumen yang memiliki *term*(df)

Tabel 2.6 Menghitung DF

Menghitung Document Frequency(df)

No	Term	Frequency
1	Saya	d1,d2,d3
2	Sangat	d1
3	Tertarik	d1,d2
4	Masuk	d1
5	Jurusan	d1
6	Hokum	d1
7	Karena	d1,d2
8	Ingin	d1,d2,d3
9	Jadi	d1,d2,d3
10	Orang	d1
11	Ahli	d1
12	Hokum	d1
13	Pengacara	d1
14	Public	d1
15	Speaking	d1
16	Bias	d1
17	Yakin	d1
18	Ilmu	d1
19	pengetahuan	d1,d3
20	Social	d1
21	Menurut	d2
22	Cukup	d2
23	bagus	d2
24	Bidang	d2
25	It	d2
26	ambil	d2
27	informatika	d2
28	komputer	d2
29	cita-cita	d2
30	bisnis	d2,d3
31	punya	d1
32	lebih	d2,d3
33	hapal	d2
34	teknik	d2
35	jaringan	d2
36	asah	d3
37	kemampuan	d3
38	dalam	d3
39	salah	d3
40	satu	d3
41	ternama	d3
42	indonesia	d3
43	pengusaha	d3
44	sendiri	d3
45	belum	d3
46	tau	d3
47	apa	d3
48	tetapi	d3
49	bisa	d3
50	hitung	d3
51	akuntansi	d3

Langkah Terakhir yaitu mengalikan TF dan IDF

Tabel 2.7 Mengalikan TFIDF

No	Term	Dokumen			idf	tf.idf		
		D1	D2	D3		D1	D2	D3
1	Saya	1	1	1	0,333333	0,33333333	0,33333333	0,333333
2	Sangat	1	0	0	1	1	0	0
3	tertarik	1	1	0	0,5	0,5	0,5	0
4	Masuk	1	0	0	1	1	0	0
5	jurusan	1	0	0	1	1	0	0
6	Hokum	1	0	0	1	1	0	0
7	Karena	1	1	0	0,5	0,5	0,5	0
8	Ingin	1	1	1	0,333333	0,33333333	0,33333333	0,333333
9	Jadi	1	1	1	0,333333	0,33333333	0,33333333	0,333333
10	Orang	1	0	0	1	1	0	0
11	Ahli	1	0	0	1	1	0	0
12	Hokum	1	0	0	1	1	0	0
13	pengacara	1	0	0	1	1	0	0
14	Public	1	0	0	1	1	0	0
15	speaking	1	0	0	1	1	0	0
16	Bisa	1	0	0	1	1	0	0
17	Yakin	1	0	0	1	1	0	0
18	Ilmu	1	0	0	1	1	0	0
19	pengetahuan	1	0	1	0,5	0,5	0	0,5
20	Social	1	0	0	1	1	0	0
21	menurut	0	1	0	1	0	1	0
22	Cukup	0	1	0	1	0	1	0
23	bagus	0	1	0	1	0	1	0
24	Bidang	0	1	0	1	0	1	0
25	It	0	1	0	1	0	1	0
26	ambil	0	1	0	1	0	1	0
27	informatika	0	1	0	1	0	1	0
28	komputer	0	1	0	1	0	1	0
29	cita-cita	0	1	0	1	0	1	0
30	bisnis	0	1	1	0,5	0	0,5	0,5
31	punya	0	1	0	1	0	1	0
32	lebih	0	1	1	0,5	0	0,5	0,5
33	hapal	0	1	0	1	0	1	0
34	teknik	0	1	0	1	0	1	0
35	jaringan	0	1	0	1	0	1	0
36	asah	0	0	1	1	0	0	1

37	kemampuan	0	0	1	1	0	0	1
38	dalam	0	0	1	1	0	0	1
39	salah	0	0	1	1	0	0	1
40	satu	0	0	1	1	0	0	1
41	ternama	0	0	1	1	0	0	1
42	indonesia	0	0	1	1	0	0	1
43	pengusaha	0	0	1	1	0	0	1
44	sendiri	0	0	1	1	0	0	1
45	belum	0	0	1	1	0	0	1
46	tau	0	0	1	1	0	0	1
47	apa	0	0	1	1	0	0	1
48	tetapi	0	0	1	1	0	0	1
49	bisa	0	0	1	1	0	0	1
50	Hitung	0	0	1	1	0	0	1
51	Akuntansi	0	0	1	1	0	0	1

Setelah langkah terakhir di dapatkan dengan menentukan bobot suatu kata dengan rumus TF x IDF, maka hasil proses perhitungan disimpan dalam *database* dan akan dilanjutkan dengan tahap pengujian dengan memberikan pertanyaan yang merupakan tahapan akhir proses seperti pada table di bawah ini.

Contoh pertanyaan:
skil apa yang kamu kuasai jika ingin masuk jurusan teknik informatika?

Tabel 2.9 Pengujian Dokumen

TF	D1	D2	D3
skil	0	0	0
kuasai	0	0	0
ingin	0,333333	0,333333	0,333333
masuk	1	1	1
jurusan	1	1	1
teknik	0	1	0
informatika	0	1	0
Total	2,333333	4,333333	2,333333

Jadi pertanyaan yang paling relavan dengan jawabannya terdapat pada dokumen 2, Tabel 7 menunjukan seberapa mirip antara *query* dengan sampel teks kalimat jawaban.

2.2.7 K-Nearest Neighbor (KNN)

KNN merupakan merupakan suatu model vector dari bobot *tf.idf* yang gunanya untuk menempatkan suatu nilai numerik pada dokumen agar dapat dihitung seberaa dekat nilai antar dokumen yang artinya menghitung tingkat kemiripan antar dokumen. Metode ini untuk melakukan klasifikasi terhadap objek yang berdasarkan dari data yang jaraknya paling dekat dengan objek tersebut dengan mencari nilai jarak Euclidian dan juga mencari nilai k yang paling baik untuk mendapatkan hasil yang baik pula.

Dalam proses mencari jarak tersebut untuk mengetahui dengan menggunakan rumus di bawah ini:

$$d = | xi - yi | = \sqrt{\sum (xi - yi)\, n\, 2}$$

Keterangan : xi = Data Uji

yi = Data Latih

i = Variabel Data

d = Jarak Dalam metode K-NN

2.2.8 *Confusion Matrix*

Ialah metode yang digunakan sebagai perhitungan suatu akurasi pada pengujian akurat suatu hasil pencarian suatu evaluasi nilai *precision*, *accuracy*, *recall*, yang mana *precision* menemukan peringkat yang paling relavan dan di definisikan sebagai suatu presentase dokumen yang di ambil dan benar benar relavan terhaap *query* [2]. Rumus *confusion matrix* adalah sebagai berikut:

Tabel 2.10 Rumus *Confusion matrix*

Documen	Nilai Sebenarnya	
	Relevan	Non relevant
Retrieve d	True Positive (tp)	False Positive (fp)
Not Retrieve	False Negative (fn)	True Negative (tn)

Keterangan:

- TP (*true positive)* = jumlah suatu prediksi benar dari data yang *relevant*
- FP (*false positive)* = jumlah suatu prediksi yang salah dari data yang tidak *relevant*
- FN (*false negative)* = jumlah suatu prediksi yang salah dari data yang tidak *relevant*
- TN (*true negative)* = jumlah suatu prediksi yang benar dari data yang *relevant*

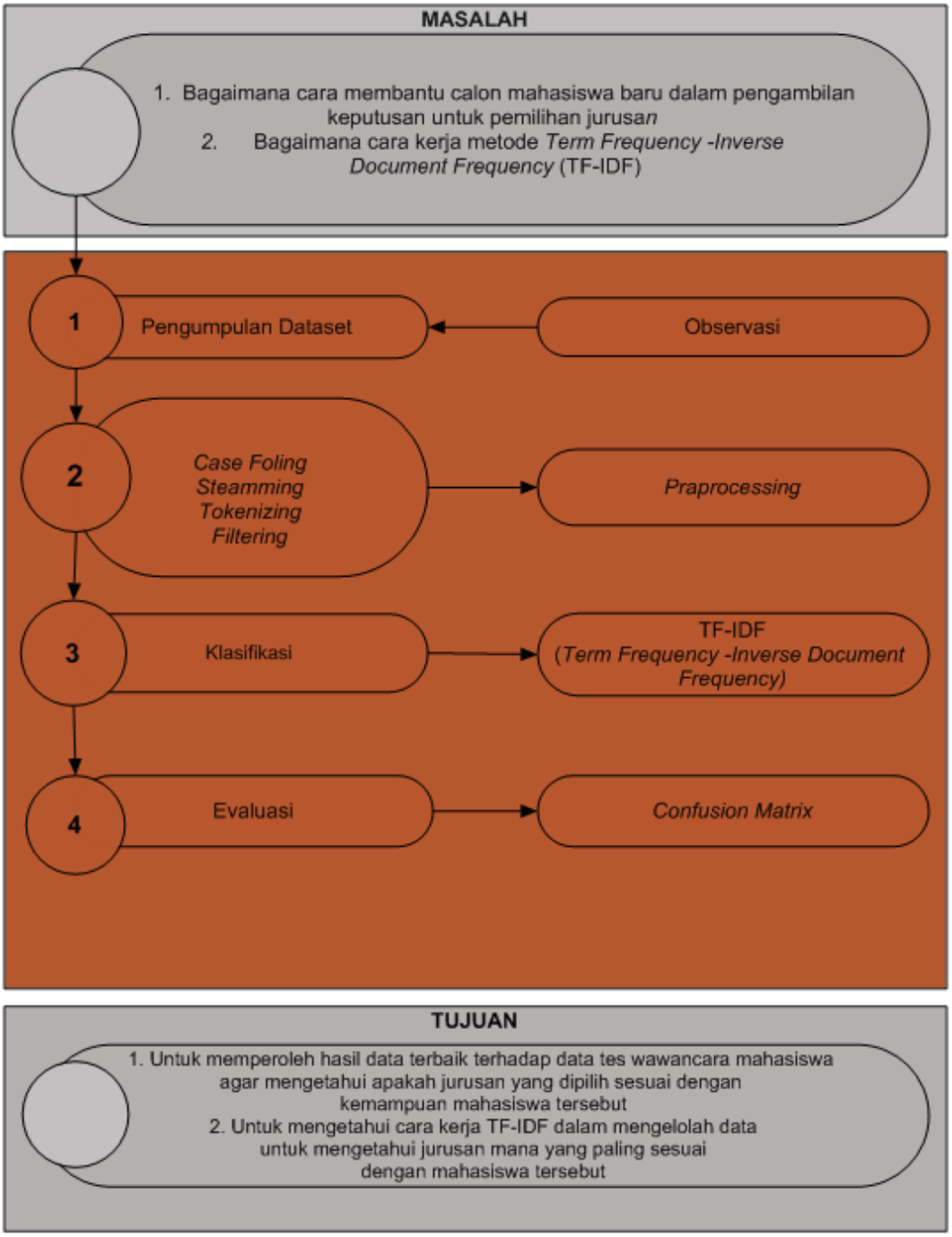
$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad (1)$$

$$Presisi = \frac{TP}{FP+TP} * 100\% \quad (2)$$

$$Recall = \frac{TP}{FN+TP} * 100\% \quad (3)$$

Gambar 2.3 Rumus *Confusion Matrix*

2.3 Kerangka Pikir



Gambar 2.5 : kerangka Pikir

BAB III

METODE PENELITIAN

3.1 Jenis, Metode, Subjek, Objek, Waktu, dan Lokasi Penelitian

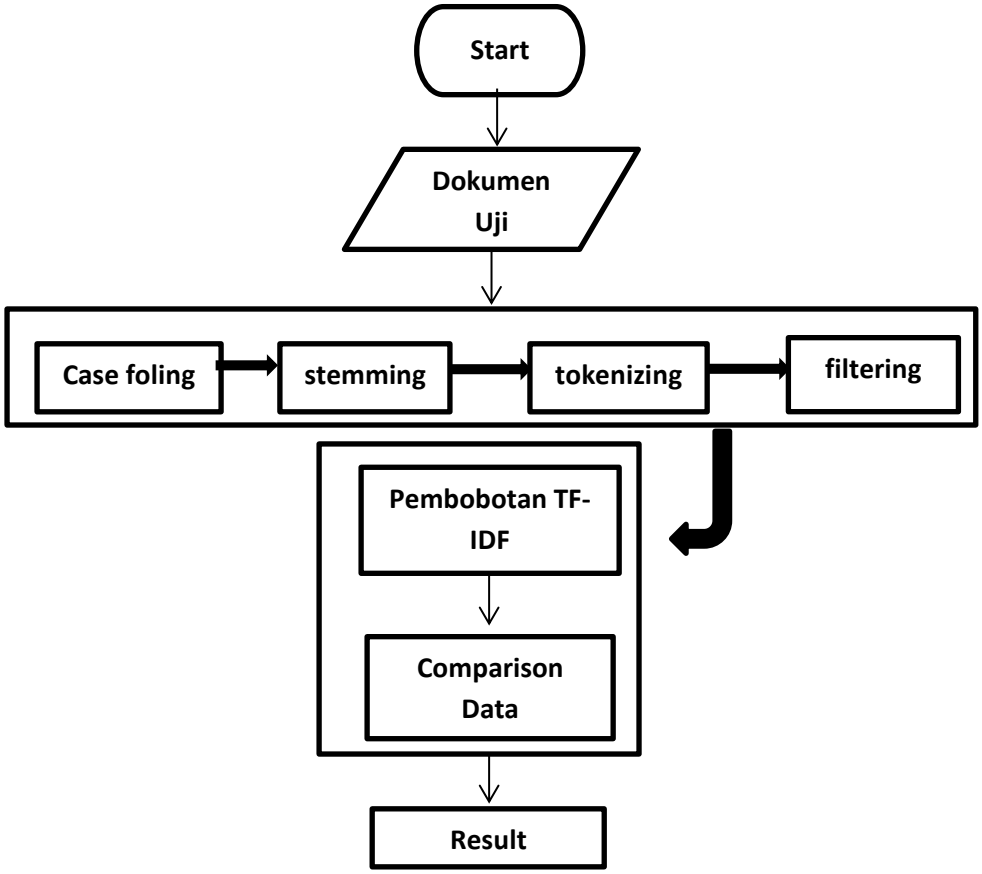
Dilihat dari penjelasannya sebelumnya, penelitian ini adalah suatu penelitian yang berfokus pada suatu data tes calon mahasiswa dalam menerapkan suatu metode agar memberikan solusi dalam suatu permasalahan agar dapat mempermudah. Dilihat dari informasi yang diolah, maka penelitian ini adalah suatu penelitian yang bersifat kuantitatif.

Dilihat dari perlakuan terhadap data, maka penelitian ini merupakan penelitian konfirmatori, karena menguji coba metode atau teori yang digunakan jika dilihat dari data yang akan diperoleh. Subjek yang digunakan pada penelitian ini adalah data tes calon mahasiswa, Objeknya ialah mahasiswa yang mengisi lembaran tes wawancara, Penelitian ini dimulai dari bulan Oktober sampai Februari dengan berlokasi pada Universitas Ichsan Gorontalo.

3.2 Dataset

Pada penelitian ini data yang akan digunakan ialah data Primer yaitu sebuah data hasil tes wawancara mahasiswa baru pada universitas ichsan gorontalo, Data yang diperoleh dari Panitia Maba tersebut adalah data yang diambil dari mahasiswa yang telah melakukan pengisian tes berupa, jurusan sebelumnya pada saat sekolah, hobi, skil, cita-cita, semua yang berhubugan dengan sesuatu yang dapat memperkuat bahwa mahasiswa tersebut bisa atau layak pada jurusan yang dipilihnya. Dan juga menggunakan data sekunder yang diperoleh dari jurnal terkait yang dengan penelitian ini.

3.3 Pemodelan



Gambar 2.1 Alur Proses

3.4 *Preprocessing*

Sebelum data diolah hal yang dilakukan terlebih dahulu yaitu mengumpulkan *dataset*, selanjutnya yaitu proses dimana melakukan tahapan *processing* data. Pada tahapan ini adalah suatu tahap yang dikerjakan sebelum melakukan proses klasifikasi, Pada tahap ini melakukan suatu pengolahan data awal menjadi data yang akan diolah ke tahap selanjutnya.

Pada tahapan *Preprocessing* terdapat beberapa hal yang akan dilakukan seperti jika terdapat huruf capital pada awalan kata maka akan diubah menjadi huruf kecil (*case folding)*, mengubah suatu kata dengan menghilangkan kata yang berimbuhan (*steaming)*, memisahka kalimat menjadi suatu kata (*tokenization)*, pada tahapan akhir yaitu memisahkan kata yang tidak penting (*filtering stopword*).

Dengan menggunakan *tools* python dengan library NLTK (*Natural Languange Tool Kit*).

3.5 Pembobotan TF-IDF (*Term Frequency-Inverse Documentd Frequency)*

Pembobotn TF-IDF (*Term Frequency – Inverse Document Frequency)* yaitu suatu proses yang melakukan perhitungan pembobotan suatu kata, frekuensi munculnya suatu kata pada dokumen tentu disebut dengan *Term Frequency* (TF) dan dokumen yang mengandung kata disebut dengan *Inverse Document Frequency* (IDF). Pada tahap pembobotan TF-IDF yaitu proses perhitungan bobot antar TF yang akan dikalikan dengan IDF untuk mengetahui seberapa benyak munculnya *term* pada setiap kata yang terdapat pada dokumen dengan proses perhitungan yang terdapat pada rumus (1).

3.6 K-Nearest Neighbor (KNN)

algoritma KNN merupakan algoritma klasifikasi yang bekerja dengan mengambil sejumlah K data terdekat (tetangganya) sebagai acuan untuk menentukan kelas dari data baru. Algoritma ini mengklasifikasikan data berdasarkan jarak terdekat atau kemiripan kedekatannya terhadap data lainnya. Dilakukan dengan memasukan nilai *term* yangdihasilkan dari klasifikasi pembobotan TF.IDF

3.7 Hasil Klasifikasi

Hasil klasifikasi pada penelitian ini adalah berupa seberapa akurat uji coba metode yang digunakan. Dari uji coba metode ini menghasilkan nilai yang diperoleh seperti kategori jurusan apa yang cocok untuk mahasiswa tersebut, serta kemiripan antar dokumen tes mahasiswa dari tiap tiap prodi dengan dokume mahasiswa baru

3.8 Evaluasi

Metode *Confusion matrix* digunakan pada tahap evaluasi yaitu sebagai metode yang dapat mengukur seberapa baik metode yang digunakan, sehingga dilakuka proses perhitungan menggunakan *confusion matrix* guna menguji apabila metode ini akurat dengan hasil klasifikasi yang diperoleh dari metode yang digunakan.

BAB IV

HASIL PENELITIAN

4.1. Pengumpulan Data

Data yang digunakan pada penelitian ini adalah suatu data yang berupa Text wawancara yang di ambil dari satu semester dalam tahun 2022/2023 sebanyak 100 data. Dari data hasil wawancara yang diperoleh berbentuk text. Selanjutnya data disimpan dalam database yang diperoleh untuk diproses pada tahapan *preprocessing*.

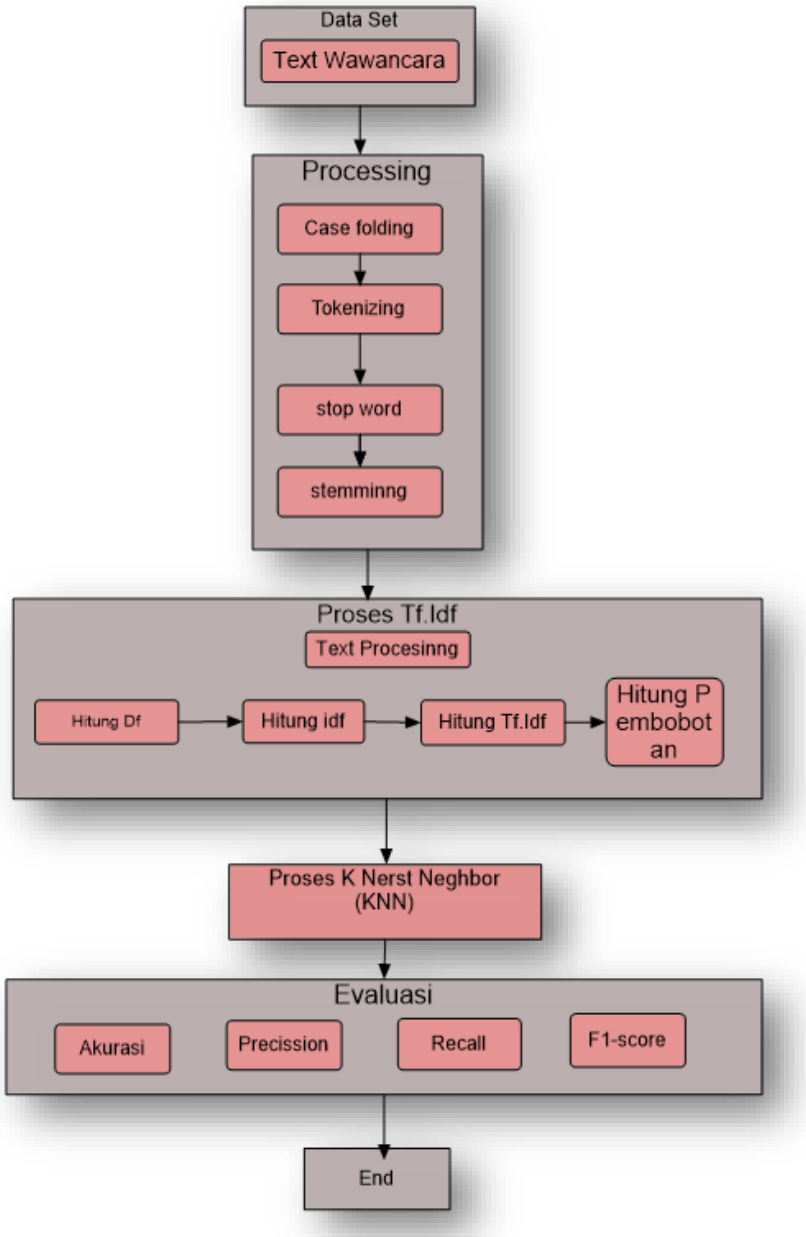
Berdasarkan dari hasil pengumpulan data, diperoleh data keseluruhan sebagai berikut :

Tabel 4.1 Hasil Pengumpulan Data

No	Nama Mahasiswa	Pertnyaan 1	Pertanyaan 2	Pertanyaan 3	Pertanyaan 4	Jurusan sekarang
1.	Farhan Hamidjun	saya tertarik karena saya ingin melanjutkan jurusan dari sekolah	cita-cita saya ingin menjadi konsultan	kelebihan saya road	saya bisa menggambar	DKV
2.	Rahmawati putri m. Razak	karena saya ingin belajar komputer	ingin menjadi programer	memiliki kelebihan tentang belajar komputer	bisa menggunakan komputer	teknik informatika
...	...	...	...	...	...	...
100	aprilia ngadi	ingin belajar tentang cara mengelolah suatu data	menjadi pengusaha	dalam jurusan saya bisa tensi, yaitu mengukur tekanan darah	mengecek tekanan darah dan bermain game	Akuntansi

4.2 Model Usulan

Dibawah inni ialah suatu model atau tahapa yang akan dilakukan untuk penelitian ini adapun *tools* atau suatu alat bantu yang di gunakan adalah menggunakan Bahasa pemograman python dengan *library streamlit* untuk menampilkan hasil pemodelann dalam bentuk *framework* seperti :



Gambar 4.1 *Flowchart*

4.3 Penerapan Metode

Berikut dibawah ini merupakan perhitungan secara manual penyelesaian menggunakan metode *term frequency-inverse document frequency* :

Tabel 4.2 Data Tes Wawancara

No	Pertanyaan 1	Pertanyaan 2	Pertanyaan 3	Pertanyaan 4	Jurusan sekarang
1.	saya tertarik karena saya ingin melanjutkan jurusan dari sekolah	cita-cita saya ingin menjadi konsultan	kelebihan saya road	saya bisa menggambar	DKV
2.	karena saya ingin belajar komputer	ingin menjadi programer	memiliki kelebihan tentang belajar komputer	bisa menggunakan komputer	teknik informatika
...	...	...	...	...	...
98	alasan saya masuk jurusan ilmu komunikasi ini yaitu ingin belajar atau melatih public speaking, dan di jurusan ilmu komunikasi ini banyak lapangan pekerjaan seperti menjadi	menjadi konten kreator dan pengusaha	kelebihan saya yaitu bernyanyi	skill yang saya kuasai yaitu berpengalaman menjadi tour guide di SMKN 1 gorontalo	ilmu komunikasi

	konten kreator atau MC di acara acara tertentu				
100	ingin belajar tentang cara mengelolah suatu data	menjadi pengusaha	dalam jurusan saya bisa tensi, yaitu mengukur tekanan darah	mengecek tekanan darah dan bermain game	akuntansi

4.2.1 *Preprocessing*

a) *Case Folding* suatu proses yang dilakukan untuk mengubah huruf besar yang berada dalam dokumen menjadi huruf kecil. Berikut dibawah ini adalah proses *Case Foldinng :*

Tabel 4.3 Hasil *Case Foldinng*

JR	Hasil *Case Folding*
J1	"saya tertarik karena saya ingin melanjutkan jurusan dari sekolah multimedia makannya saya ambil jurusan dkv cita-cita saya ingin menjadi konsultan,saya suka foto,video,apapun yang berbaur seni, kelebihan saya road, saya bisa menggambar"
J2	"karena saya ingin belajar komputer,ingin menjadi programer,memiliki kelebihan tentang belajar komputer,bisa menggunakan komputer"
J3	"saya tertarik karena saya ingin jadi orang teknik dan rencana saya maau buka servis di papua,teknisi,rakit hardware,olahraga"
J4	"sesuai jurusan di sekolah ekonomi akuntansi ,menjadi auditing,berhitung"

J5	"sesuai jurusan di sekolah ekonomi akuntansi ,menjadi auditing,berhitung"
J6	"karena dengan keinginan saya sendiri karena nantinya setelah dari sini saya akan melanjutkan masuk polri,ingin menjadi anggota polri,yaitu sepak bola, olahraga,sepak bola"
J7	"karena saya ingin memperdalam tentang ilmu pertanian,cita-cita ingin memiliki kebun"
J8	"karena ketertarikan dan mempunyai bakat,ingin jadi pelukis,menggambar,menggambar"
J9	"karena saya ingin mempelajari cara berkomunikasi dengan baik,prisenter,berkomunikasi dengan baik,bisa berinteraksi dengan baik"
J10	"saya tertarik dan kemauan saya sendiri dan bisa menambah wawasan, bisa mencari tau apa itu jurusan sistem informasi,cita-cita saya ingin jadi pengusaha,kelebihan saya bisa menjual makanan bisa membantu orang tua saya,saya sekarang lagi belajar komputer"
J11	"ada basic karena sering mengikuti keluarga bekerja dibidang listrik,menejer pln,olahraga,olahraga dan sepak bola"
J12	"ingin belajar terkait dengan pembuatan alat alat peranian,petani yang sukses, mananam tanaman"
J13	"saya punya kebun dan sawah,ingin membangun bisnis yang ada kaitannya dengan pertanian,suka membaca dan main bola volly"
…	…
40	"karena suka hukum,jadi pembisnis dan pengacara,bermain bola"

b) *Tokenizing* ialah suatu proses yang dilakukan untuk memisahkan dokumen dari suatu kalimat menjadi kumpulan bagian – bagian kata tertentu, dibawah ini adalah proses tokenizing :

Tabel 4.4 Hasil *Tokenizinng*

JR	Hasil *Tokenizing*
J1	"saya","tertarik","karena","saya","ingin","melanjutkan","jurusan","dari","sekolah","multimedia","makannya","saya","ambil","jurusan","dkv","cita","cita","saya","ingin","menjadi","konsultan","saya","suka","foto","video","apapun","yang","berbaur","seni","kelebihan","saya","road","saya","bisa","menggambar"
J2	"karena","saya","ingin","belajar","31okum31er","ingin","menjadi","31okum31er3131","memiliki","kelebihan","tentang","belajar","31okum31er","bisa","menggunakan","komputer"
J3	"saya","tertarik","karena","saya,"ingin","jadi","orang","31okum31","dan","rencana","saya","maau","buka","servis","di","papua","teknisi","rakit","hardware","olahraga"
J4	"sesuai","jurusan","di","sekolah","ekonomi","akuntansi","menjadi","auditing","berhitung"
J5	"31okum31e","saya","memilih","jurusan","ini","karena","saya","ingin","lebih","dalam","untuk","memperdalam","jurusan","ekonomi","manajeme","pengusaha","pramuka"

J6	"karena","dengan","keinginan","saya","sendiri","karena","nantinya","setelah","dari","sini","saya","akan","melanjutkan","masuk","polri","ingin","menjadi","anggota","polri","yaitu","sepak","bola","olahraga","sepak","bola"
J7	"karena",saya","ingin","memperdalam","tentang","ilmu","pertanian","cita","cita","ingin","memiliki","kebun"
J8	"karena","ketertarikan","dan","mempunyai","bakat","ingin","jadi","pelukis","menggambar","menggambar"
J9	"karena","saya","ingin","mempelajari","cara","berkomunikasi","dengan","baik","32okum32er32","berkomunikasi","dengan","baik","bisa","berinteraksi","dengan","baik"
J10	"saya","tertarik","dan","kemauan","saya,"sendiri","dan,"bisa","menambah","wawasan","bisa","mencari","tau","apa","itu","jurusan","32okum32","informasi","cita","cita","saya","ingin","jadi","pengusaha","kelebihan","saya","bisa","menjual","makanan","bisa","membantu","orang","tua","saya,"saya","sekarang","lagi","belajar","32okum32er"
J11	"ada","basic","karena","sering","mengikuti","keluarga","bekerja","dibidang","listrik","menejer","pln","olahraga","olahraga","dan","sepak","bola"
J12	"ingin","belajar","terkait","dengan","pembuatan","alat","alat","peranian","petani","yang","sukses","mananam","tanaman"
J13	"saya","punya","kebun","dan","sawah","ingin","membangun","bisnis","yang","ada","kaitannya","dengan","p

	ertanian”,”suka”,”membaca”,”dan”,”main”,”bola”,”vol ly”
…	…
J40	“karena”,”suka”,”33okum”,”jadi”,”pembisnis”,”dan”,” pengacra”,”bermain”,”bola”

c) *Stopword removal* ialah suatu tahapan yang dimana membuang kata sehingga menyisakan kata yang penting dibawah ini adalah proses *stopword removal* :

Tabel 4.5 Hasil *Stopword removal*

JR	Hasil *Tokenizing*
J1	"tertarik","melanjutkan","jurusan","sekolah","multi media","makannya","ambil","jurusan","dkv","cita","c ita","menjadi","konsultan","suka","foto","video","a papun","berbaur","seni","kelebihan","road","menggam bar"
J2	"belajar","komputer","menjadi","programer","memilik i","kelebihan","belajar","komputer","menggunakan"," komputer"
J3	"tertarik","jadi","orang","teknik","rencana","maau" ,"buka","servis","papua","teknisi","rakit","hardwar e","olahraga"
J4	"sesuai","jurusan","sekolah","ekonomi","akuntansi", "menjadi","auditing","berhitung"

J5	"alasan","memilih","jurusan","lebih","memperdalam","jurusan","ekonomi","manajeme","pengusaha","pramuka"
J6	"keinginan","sendiri","nantinya","sini","melanjutkan","masuk","polri","menjadi","anggota","polri","sepak","bola","olahraga","sepak","bola"
J7	"memperdalam","ilmu","pertanian","cita","cita","memiliki","kebun"
J8	"ketertarikan","mempunyai","bakat","jadi","pelukis","menggambar","menggambar"
J9	"mempelajari","cara","berkomunikasi","baik","presenter","berkomunikasi","baik","berinteraksi","baik"
J10	"tertarik","kemauan","sendiri","menambah","wawasan","mencari","tau","apa","jurusan","sistem","informasi","cita","cita","jadi","pengusaha","kelebihan","menjual","makanan","membantu","orang","tua","sekarang","belajar","komputer"
J11	"basic","sering","mengikuti","keluarga","bekerja","dibidang","listrik","menejer","pln","olahraga","olahraga","sepak","bola"
J12	"belajar","terkait","pembuatan","alat","alat","peranian","petani","sukses","mananam","tanaman"
J13	"punya","kebun","sawah","membangun","bisnis","kaitannya","pertanian","suka","membaca","main","bola","volly"
…	…
J40	"suka","hukum","jadi","pembisnis","pengacra","bermain","bola"

a) *Stemming* ialah tahapan yang dimana mengubah kata yang berimbuhan menjadi suatu kata dasar dibawah ini adalah 10 data untuk proses *stemming*

Tabel 4.5 Hasil *Stemming*

JR	Hasil *Stemming*
J1	"tarik","lanjut","jurus","sekolah","multimedia","makan","ambil","jurus","dkv","cita","cita","jadi","konsultan","suka","foto","video","apa","baur","seni","lebih","road","gambar"
J2	"ajar","komputer","jadi""programer","milik","lebih","ajar","komputer","guna","komputer"
J3	"tarik","jadi","orang","teknik","rencana","maau","buka","servis","papua","teknisi","rakit","hardware","olahraga"
J4	"sesuai","jurus","sekolah","ekonomi","akuntansi","jadi","auditing","hitung"
J5	"alas","pilih","jurus","lebih","dalam","jurus","ekonomi","manajeme","usaha","pramuka"
J6	"ingin","sendiri","nanti","sini","lanjut","masuk","polri","jadi","anggota","polri","sepak","bola","olahraga","sepak","bola"
J7	"dalam","ilmu","tani","cita","cita","milik","kebun"
J8	"tari","punya","bakat","jadi","peluk","gambar","gambar"
J9	"ajar","cara","komunikasi","baik","prisenter","komunikasi","baik","interaksi","baik"

J10	"tarik","mau","sendiri","tambah","wawas","cari","tau","apa","jurus","sistem","informasi","cita","cita","jadi","usaha","lebih","jual","makan","bantu","orang","tua","sekarang","ajar","komputer"
J11	"basic","sering","ikut","keluarga","kerja","bidang","listrik","menejer","pln","olahraga","olahraga","sepak","bola"
J12	"ajar","kait","buat","alat","alat","ani","tani","sukses","mananam","tanam"
J13	"punya","kebun","sawah","bangun","bisnis","kait","tani","suka","baca","main","bola","volly"
…	…
40	"suka","hukum","jadi","bisnis","pengacra","main","bola"

Pada Tahap selanjutnya adalah suatu proses dimana menghitung berapa banyak kata sering muncul atau digunakan mahasiswa pada setiap Teks wawancara yang telah di isi, dibawah ini adalah sebagian proses menghitung kemunculan kata pada TF (*Term Frequency*) :

Tabel 4.7 Menghitung Kemunculan Kata

Term	J1	J2	J3	J4	J5	J6	J7	J39	J40	Q
Tarik	1	0	1	0	0	0	…	0	0	0
Lanjut	1	0	0	0	0	1	…	0	0	0
Jurus	1	0	0	1	1	0	…	0	0	1
Sekolah	1	0	0	1	0	0	…	0	0	0
multimedia	1	0	0	0	0	0	…	0	0	0
Makan	1	0	0	0	0	0	…	0	0	0
Ambil	1	0	0	0	0	0	…	0	0	0
Dkv	1	0	0	0	0	0	…	0	0	0

Cita	1	0	0	0	0	0	…	0	0	0
Jadi	1	1	1	1	0	0	…	0	1	0
konsultan	1	0	0	0	0	0	…	0	0	0
Suka	1	0	0	0	0	0	…	0	1	0
Foto	1	0	0	0	0	0	…	0	0	0
Video	1	0	0	0	0	0	…	0	0	0
Apa	1	0	0	0	0	0	…	0	0	0
Baur	1	0	0	0	0	0	…	0	0	0
Seni	1	0	0	0	0	0	…	0	0	0
Lebih	1	0	0	0	1	0	…	0	0	0
Road	1	0	0	0	0	0	…	0	0	0
Gambar	1	0	0	0	0	0	…	0	0	0
Ajar	0	1	0	0	0	0	…	0	0	1
komputer	0	1	0	0	0	0	…	1	0	1
programer	0	1	0	0	0	0	…	1	0	0
Milik	0	1	0	0	0	0	…	0	0	0
Lebih	0	1	0	0	0	0	…	0	0	0
Orang	0	0	1	0	0	0	…	0	0	0
Teknik	0	0	1	0	0	0	…	0	0	1
Rencana	0	0	1	0	0	0	…	0	0	0
Mau	0	0	1	0	0	0	…	0	0	0
Buka	0	0	1	0	0	0	…	0	0	0
Servis	0	0	1	0	0	0	…	0	0	0
Papua	0	0	1	0	0	0	…	0	0	0
Teknisi	0	0	1	0	0	0	…	0	0	0
Rakit	0	0	1	0	0	0	…	0	0	0
hardware	0	0	1	0	0	0	…	0	0	0
Olahraga	0	0	1	0	0	1	…	0	0	0
Sesuai	0	0	0	1	0	0	…	0	0	0
Ekonomi	0	0	0	1	1	0	…	0	0	0
akuntansi	0	0	0	1	0	0	…	0	0	0
Auditing	0	0	0	1	0	0	…	0	0	0
Hitung	0	0	0	1	0	0	…	0	0	0
Alas	0	0	0	0	1	0	…	0	0	0
Pilih	0	0	0	0	1	0	…	0	0	0
Dalam	0	0	0	0	1	0	…	0	0	0
manajemen	0	0	0	0	1	0	…	1	0	0
Usaha	0	0	0	0	1	0	…	1	0	0
Pramuka	0	0	0	0	1	0	…	0	0	0
…	…	…	…	…	…	…	…	0	0	0
pengacara	0	0	0	0	0	0	0	0	0	0

4.2.2 Pembobotan TF.IDF

Setelah menghitung kemunculan kata dari setiap dokumen, selanjutnya dilakukan proses perhitungan TF.IDF (*Term Frequency-Inverse Document Frequency*). Proses perhitungan TF.IDF ini dapat dilihat pada tabel dibawah ini.

Tabel 4.7 Perhitungan DF(*Document Frequency*)

Term	Df		Term	df		Term	df
tarik	9		Diri	3		voli	2
lanjut	3		Nanti	1		saya	1
jurus	12		Sini	1		game	4
sekolah	3		Masuk	3		jaringan	4
multimedia	1		Polri	1		informatika	3
makan	2		Sepak	2		it	1
ambil	1		Bola	5		dua	1
dkv	1		Ilmu	5		edit	2
cita	6		Tani	3		reporter	2
jadi	17		Kebun	2		polisi	3
konsultan	1		Tari	1		sedikit	1
suka	6		Punya	4		prospek	1
foto	2		Bakat	1		depan	1
video	1		Peluk	1		sangat	2
apa	2		Cara	3		janji	1
baur	1		komunikasi	2		lapang	1
seni	1		Baik	2		bahagia	1
lebih	7		prisenter	3		sosial	1
road	1		interaksi	2		mampu	2
gambar	4		tambah	2		percaya	1
ajar	12		Cari	1		elektro	1
komputer	13		Tau	5		hukum	3
programer	3		System	1		teknologi	1
milik	5		informasi	2		tinggi	1
lebih	1		Jual	1		aplikasi	2
orang	6		Bantu	1		peluang	1
teknik	4		Tua	4		jaksa	2
rencana	2		sekarang	1		daya	1
mau	7		Basic	1		tangkap	1
buka	1		Sering	1		paham	2
servis	1		Ikut	1		cukup	1

papua	1		keluarga	4		publik	2
teknisi	1		Kerja	6		speaking	2
rakit	1		Bidang	7		sampai	1
hardware	2		Listrik	2		arsitek	1
olahraga	3		manajer	1		hewan	1
sesuai	1		Pln	1		rumah	1
ekonomi	3		Kait	3		uang	1
akuntansi	4		Buat	7		debat	1
auditing	1		Alat	2		acara	1
hitung	3		Sukses	1		tv	1
alas	3		menanam	1		bahas	1
pilih	4		tanaman	1		hafal	1
dalam	7		Sawah	1		wartawan	1
manajemen	4		Bangun	2		data	1
usaha	10		Bisnis	3		instalasi	1
pramuka	1		Baca	1		pengacara	1
ingin	2		Main	5			

Dari tabel diatas adalah sebagian dari kata yang diambil dari proses perhitungan DF (*Document Frequency*) yang dimana menghitung kata yang muncul dari setiap dokumen. Setelah dilakukan perhitungan DF, maka dilakukan proses perhitungan IDF secara manual dengan menggunakan rumus dibawah ini:

$$idf_t = \log\left(\frac{n}{df_t}\right)$$

Term [tarik] :

$$idf_t = \log\left(\frac{n}{df_t}\right)$$

$$idf_t = \log\left(\frac{40}{9}\right)$$

$$= 4{,}44$$

Term [lanjut] :

$$idf_t = \log\left(\frac{n}{df_t}\right)$$

$$idf_t = \log\left(\frac{40}{3}\right)$$

$$= 13{,}33$$

Term [jurus] :

$$idf_t = \log\left(\frac{n}{df_t}\right)$$

$$idf_t = \log\left(\frac{40}{12}\right)$$

$$= 3{,}33$$

Term [sekolah] :

$$idf_t = \log\left(\frac{n}{df_t}\right)$$

$$idf_t = \log\left(\frac{40}{3}\right)$$

$$= 13{,}33$$

Term [Multimedia] :

$$idf_t = \log\left(\frac{n}{df_t}\right)$$

$$idf_t = \log\left(\frac{40}{1}\right)$$

$$= 40$$

Lakukan perhitungan sampai dengan kata terakhir yang ada pada table di atas

Selanjutnya hasil dari perhitungan IDF (*Inverse Document Frequency*) dapat dilihat pada tabel dibawah ini :

Tabel 4.9 Perhitungan IDF
(*Inverse Document Frequemcy*)

Term	Df	df log(n/df)		Term	Df	df log(n/df)		Term	Df	df log(n/df)
tarik	9	4,44444444		Diri	3	13,3333333		voli	2	20
lanjut	3	13,3333333		Nanti	1	40		game	4	10
jurus	12	3,33333333		Sini	1	40		jaringan	4	10
sekolah	3	13,3333333		Masuk	3	13,3333333		informatika	3	13,3333333
multimedia	1	40		Polri	1	40		it	1	40
makan	2	20		Sepak	2	20		dua	1	40
ambil	1	40		Bola	5	8		edit	2	20
dkv	1	40		Ilmu	5	8		reporter	2	20
cita	6	6,66666667		Tani	3	13,3333333		polisi	3	13,3333333
jadi	17	2,35294118		Kebun	2	20		sedikit	1	40
konsultan	1	40		Tari	1	40		prospek	1	40
suka	6	6,66666667		Punya	4	10		depan	1	40
foto	2	20		Bakat	1	40		sangat	2	20
video	1	40		Peluk	1	40		janji	1	40
apa	2	20		Cara	3	13,3333333		lapang	1	40
baur	1	40		Komunikasi	2	20		bahagia	1	40
seni	1	40		Baik	2	20		sosial	1	40
lebih	7	5,71428571		Presenter	3	13,3333333		mampu	2	20
road	1	40		Interaksi	2	20		percaya	1	40
gambar	4	10		Tambah	2	20		elektro	1	40
ajar	12	3,33333333		Cari	1	40		hukum	3	13,3333333
komputer	13	3,07692308		Tau	5	8		teknologi	1	40

programer	3	13,3333333		System	1	40		tinggi	1	40
milik	5	8		Informasi	2	20		aplikasi	2	20
lebih	1	40		Jual	1	40		peluang	1	40
orang	6	6,66666667		Bantu	1	40		jaksa	2	20
teknik	4	10		Tua	4	10		daya	1	40
rencana	2	20		Sekarang	1	40		tangkap	1	40
mau	7	5,71428571		Basic	1	40		paham	2	20
buka	1	40		Sering	1	40		cukup	1	40
servis	1	40		Ikut	1	40		publik	2	20
papua	1	40		Keluarga	4	10		speaking	2	20
teknisi	1	40		Kerja	6	6,666666 67		sampai	1	40
rakit	1	40		Bidang	7	5,714285 71		arsitek	1	40
hardware	2	20		Listrik	2	20		hewan	1	40
olahraga	3	13,3333333		Manajer	1	40		rumah	1	40
sesuai	1	40		Pln	1	40		uang	1	40
ekonomi	3	13,3333333		Kait	3	13,333333 33		debat	1	40
akuntansi	4	10		Buat	7	5,714285 71		acara	1	40
auditing	1	40		Alat	2	20		tv	1	40
hitung	3	13,3333333		Sukses	1	40		bahas	1	40
alas	3	13,3333333		Menanam	1	40		hafal	1	40
pilih	4	10		Tanaman	1	40		wartawan	1	40
dalam	7	5,71428571		Sawah	1	40		data	1	40
manajemen	4	10		Bangun	2	20		instalasi	1	40
usaha	10	4		Bisnis	3	13,333333 33		pengacara	1	40
pramuka	1	40		Baca	1	40				
ingin	2	20		Main	5	8				

Setelah dilakukan proses perhitungan IDF (*Inverse Document Frequency*), maka dilakukan proses perhitungan selanjutnya adalah menghitung nilai TF.IDF dengan cara nilai TF × IDF sehingga menghasilkan nilai TF.IDF. Dibawah ini ialah sebagian proses perhitungan manual dengan seperti :

$$W_{dt} = TF_{dt} X\ idf_t$$

Term [tarik] : $W_{dt} = TF_{dt} X\ idf_t$

$W = 1 \times 4{,}33 = 4{,}33$

Term [lanjut] : $W_{dt} = TF_{dt} X\ idf_t$

$W = 1 \times 6{,}5 = 6{,}5$

Term [jurus] : $W_{dt} = TF_{dt} X\ idf_t$

$W = 1 \times 2{,}6 = 2{,}6$

Term [sekolah] : $W_{dt} = TF_{dt} X\ idf_t$

$W = 1 \times 6{,}5 = 6{,}5$

Term [multimedia] : $W_{dt} = TF_{dt} X\ idf_t$

$W = 1 \times 13 = 13$

Term [makan] : $W_{dt} = TF_{dt} X\ idf_t$

$W = 1 \times 6{,}5 = 6{,}5$

Term [ambil] : $W_{dt} = TF_{dt} X\ idf_t$

$W = 0 \times 13 = 0$

Term [dkv] : $W_{dt} = TF_{dt} X\ idf_t$

$W = 1 \times 13 = 13$

Lakuka perhitungan sampai dengan kata terakhir yang ada pada table di atas.

Pada tabel dibawah ini dapat dilihat keseluruhan perhitungan dari setiap dokumen seperti dibawah ini :

Tabel 4.10 Hasil Perhitungan TF.IDF

TF.IDF								
J1	J2	J3	J4	J5	J6	J7	J40	Q
4,444444	0	4,444444	0	0	0	...	0	0
13,33333	0	0	0	0	13,33333	...	0	0
3,333333	0	0	3,333333	3,333333	0	...	0	3,333333
13,33333	0	0	13,33333	0	0	...	0	0
40	0	0	0	0	0	...	0	0
20	0	0	0	0	0	...	0	0
40	0	0	0	0	0	...	0	0
40	0	0	0	0	0	...	0	0
6,666667	0	0	0	0	0	...	0	0
2,352941	2,352941	2,352941	2,352941	0	0	...	2,352941	0
40	0	0	0	0	0	...	0	0
6,666667	0	0	0	0	0	...	6,666667	0
20	0	0	0	0	0	...	0	0
40	0	0	0	0	0	...	0	0
20	0	0	0	0	0	...	0	0
40	0	0	0	0	0	...	0	0
40	0	0	0	0	0	...	0	0
5,714286	0	0	0	5,714286	0	...	0	0
40	0	0	0	0	0	...	0	0
10	0	0	0	0	0	...	0	0
0	3,333333	0	0	0	0	...	0	3,333333
0	3,076923	0	0	0	0	...	0	3,076923
0	13,33333	0	0	0	0	...	0	0
0	8	0	0	0	0	...	0	0
0	40	0	0	0	0	...	0	0
0	0	6,666667	0	0	0	...	0	0
0	0	10	0	0	0	...	0	10
0	0	20	0	0	0	...	0	0
0	0	5,714286	0	0	0	...	0	0
0	0	40	0	0	0	...	0	0
0	0	40	0	0	0	...	0	0
0	0	40	0	0	0	...	0	0
0	0	40	0	0	0	...	0	0
0	0	40	0	0	0	...	0	0
0	0	20	0	0	0	...	0	0

	0	0	13,33333	0	0	13,33333	…	0	0
	0	0	0	40	0	0	…	0	0
	0	0	0	13,33333	13,33333	0	…	0	0
	0	0	0	10	0	0	…	0	0
Total	445,845	70,09653	282,5117	135,6863	105,4286	221,3333	…	78,35294	113,0769

4.2.3 Perhitungan *KNN*

Setelah nilai TF.IDF diketahui maka, tahapan selanjutnya ialah proses perhitungan *KNN* yang artinya menghitung kesamaan dalam membandingkan kemiripan antar data/dokumen. Dengan menggunakan rumus seperti dibawah ini :

$$\boldsymbol{D(X,Y)} = \sqrt{\sum_{\boldsymbol{k}}^{\boldsymbol{n}} = (\boldsymbol{Xk} - \boldsymbol{Yk})^{\,2}}$$

Ket :

D = Jarak antara dua titik x dan y

X = Data uji

Y = Sampel data

n = Dimensi data

tahapan perhitungan secara manual untuk algoritma *KNN* dengan menggunakan rumus diatas seperti :

- Pada tahapann awal KNN ialah menentukan nilai K,
- Selanjutnya Menghitung jarak data baru di setiap label data (jarak Euclidean) dengan jarak semua data training. Untuk menghitung tingkat kesamaan dalam dokumen menggunakan rumus di atas dengan perhitungan seperti dibawah ini :

$$\text{Jarak} = \sqrt{(Mhsbaru - Mhs1)^{\,2}}$$

- Kemudian urutkan hasil Euclidean distance berdasarkan jarak K yang ditentukan jika K=3 artinya akan dipilih 3 jarak terkecil dari hasil Euclidean distance.

Dibawah ini adalah tahapan perhitungan manual KNN dengan beberapa jarak K yang di tentukan

Jarak 1 = $\sqrt{(Mhsbaru - Jr1)}^2$
= $\sqrt{(113{,}0769 - 445{,}845)}^2$
= $\sqrt{(-332{,}768)}^2 = (-332{,}768 \times -332{,}768)$
= $\sqrt{110734{,}6}$ = 332,76

Jarak 2 = $\sqrt{(Mhsbaru - Jr2)}^2$
= $\sqrt{(113{,}0769 - 70{,}09653)}^2$
= $\sqrt{(42{,}98037)}^2 = (42{,}98037 \times 42{,}98037$
= $\sqrt{1847{,}312}$ = 42,98

Jarak 3= $\sqrt{(Mhsbaru - Jr3)}^2$
= $\sqrt{(113{,}0769 - 282{,}5117}^2$
= $\sqrt{(-169{,}435)}^2 = (-169{,}435 \times -169{,}435)$
= $\sqrt{28708{,}15}$ = 169,43

Jarak 4 = $\sqrt{(Mhsbaru - Jr4)}^2$
= $\sqrt{(113{,}0769 - 135{,}6863)}^2$
= $\sqrt{(-22{,}6094)}^2 = (-22{,}6094 \times -22{,}6094)$
= $\sqrt{511{,}185}$ = 22,60

Jarak 5 = $\sqrt{(Mhsbaru - Jr5)}^2$
= $\sqrt{(113{,}0769 - 105{,}4286)}^2$
= $\sqrt{(7{,}6483)}^2 = (7{,}6483 \times 7{,}6483)$
= $\sqrt{58{,}49}$ = 7,64

Jarak 6 = $\sqrt{(Mhsbaru - Jr6)}^2$

= $\sqrt{(113{,}0769 - 221{,}3333)}^2$

=$\sqrt{(-108{,}256)}^2 = (-108{,}256 - \times 108{,}256)$

= $\sqrt{11719{,}45}$ = 7108,25

Jarak 7 = $\sqrt{(Mhsbaru - Jr7)}^2$

= $\sqrt{(113{,}0769 - 61{,}71429}^2$

=$\sqrt{(51{,}36261}^2 = (51{,}3626 \times 56{,}36261)$

= $\sqrt{2638{,}118}$ = 51,36

Jarak 8 = $\sqrt{(Mhsbaru - Jr8)}^2$

= $\sqrt{(113{,}0769 - 142{,}3529)}^2$

= $\sqrt{(-29{,}276)}^2 = (-29{,}276 \times -29{,}276)$

= $\sqrt{857{,}0842}$ = 29,27

Jarak 9= $\sqrt{(Mhsbaru - Jr9)}^2$

= $\sqrt{(113{,}0769 - 90}^2$

= $\sqrt{(23{,}0769)}^2 = (23{,}0769 \times 23{,}0769)$

= $\sqrt{532{,}5433}$ = 23,07

Jarak 10 = $\sqrt{(Mhsbaru - Jr10)}^2$

= $\sqrt{(113{,}0769 - 336{,}6362)}^2$

= $\sqrt{(-223{,}559)}^2 = (-223{,}559 \times -223{,}559$

= $\sqrt{110734{,}6}$ = 2332,76

Lakukan perhitungan sampai dengan Dokumen terakhir yang ada pada table di atas

- Kemudian urutkan hasil Euclidean distance berdasarkan jarak K yang ditentukan artinya akan dipilih jarak terkecil dari hasil Euclidean distance.

Tabel 4.12 Pengurutan Tingkat Kemiripan

D	Euclidean distance 113,0769	Urutan jarak dengan data baru	Kategori jurusan (Class)	D	Euclidean distance 113,0769	Urutan jarak dengan data baru	Kategori jurusan (Class)
Jr1	332,76		Dkv	Jr21	9,68		teknik elektro
Jr2	42,98		teknik informatika	Jr22	49,29		hukum
Jr3	169,43		teknik elektro	Jr23	55,23		manajemen
Jr4	22,6		Akuntansi	Jr24	96,28	3	teknik informatika
Jr5	7,64		Manajemen	Jr25	33,84		sistem informasi
Jr6	108,25	2	Hokum	Jr26	12,76		akuntansi
Jr7	51,36		Agroteknologi	Jr27	68,95	8	sistem informasi
Jr8	29,27		teknik arsitektur	Jr28	27,2		sistem informasi
Jr9	23,07		ilmu komunikasi	Jr29	111,65	1	hukum
Jr10	223,55		sistem informasi	Jr30	17,58		akuntansi
Jr11	190,63		teknik elektro	Jr31	80,38	6	teknik arsitektur
Jr12	62,63	9	Thp	Jr32	70,41	7	teknik informatika
Jr13	91,58	4	Agribisis	Jr33	57,33	10	teknik informatika
Jr14	22,63		Dkv	Jr34	36,66		teknik informatika
Jr15	55,87		teknik informatika	Jr35	82,41	5	akuntansi
Jr16	49,48		Manajemen	Jr36	20,59		akuntansi
Jr17	23,07		ilmu komunikasi	Jr37	126,92		hukum
Jr18	151,68		Dkv	Jr38	7,26		sistem informasi
Jr19	50,66		teknik informatika	Jr39	29,04		manajemen
Jr20	9,65		Manajemen	Jr40	34,72		hukum

Ambil sejumlah K data dengan jarak terdekat, kemudian tentukan kelas dari data baru tersebut, dibawah ini kita akan mencoba menentukan nilai K = 3 sampai 10.Untuk menentukan nilai K yang benar, Algoritma K-nearest neighbor dapat beberapa kali dicoba dengan memakai nilai K yang berbeda-beda. Hal ini bertujuan untuk mendapatkan hasil yang akurat. Penentuan nilai K ini tidak ada rumus pastinya.

Untuk Nilai K = 3 Jarak data lama dengan data baru menunjukan bahwa data baru tersebut lebih dekat jaraknya dengan kategori jurusan ilmu Hukum dan jarak ketiganya dengan jurusan Teknik informatika, dengan kelas tetangga terbanyak ilmu Hukum

K = 10 Jarak data lama dengan data baru menunjukan bahwa data baru dengan kelas tetangga terbanyak Teknik informatika 3

Data	Jarak dengan data baru	Urutan Jarak	Kategori Jurusan (Class)
Jr33	57,33	1	Teknik informatika
Jr12	62,63	2	thp
Jr27	68,95	3	System informssi
Jr32	70,41	4	Teknik informatika
Jr31	80,83	5	Teknik arsitektur
Jr35	82,41	6	Akuntansi
Jr13	91,58	7	Agrisbisnis
Jr24	96,28	8	Teknik informatika
Jr6	108,25	9	Hokum
Jr29	111,65	10	Hokum

BAB V

HASIL EVALUASI

5.1 Evaluasi Model

Setelah Model selesai dibangun maka selanjutnya yaitu melakukan evaluasi pada model tersebut, Langkah Pertama yang digunakan untuk mengevakuasi model adalah menentukan data test dari dataset yang ada dengan menggunakan :

```
 X_train, X_test, y_train, y_test = train_test_split(
v_data, dnb['label'], test_size=0.2, random_state=0)
```

Dari 100 Dataset yang ada dan dengan menggunakan persentase trainingset dan testing set =0,1%, maka di dapatkan data testing sebanyak 10 data.adapun kelas Aktual dari data test yang akan di evaluasi adalah sebagai berikut :

Tabel 5. 1 Kelas Aktual

No.	No. ID	Kelas Aktual
1	26	teknik arsitektur
2	86	teknik informatika
3	2	teknik elektro
4	55	ilmu pemerintahan
5	75	ilmu hokum
6	93	ilmu hokum
7	16	ilmu hokum
8	73	teknik informatika
9	54	teknik elektro
10	95	teknik informatika

Selanjutnya akan di lakukan beberapa kali uji coba dengan menggunakan nilai K yang berbeda beda. Karena Pada Penelitian ini kelas yang digunakan adalah Ganjil Maka Nilai K yang akan di gunakan adalah Genap

a. Uji Coba 1 (Nilai K =2)

Tabel 5.2 Hasil Uji Coba (K=2)

ID	Kelas Aktual	Kelas Prediksi
26	teknik arsitektur	ilmu komunikasi
86	teknik informatika	teknik informatika
2	teknik elektro	ilmu hukum
55	ilmu pemerintahan	ilmu komunikasi
75	ilmu hokum	agroteknologi
93	ilmu hokum	ilmu hukum
16	ilmu hokum	ilmu hukum
73	teknik informatika	ilmu hukum
54	teknik elektro	Sistem Informasi
95	teknik informatika	Sistem Informasi

Tabel 5.3 Confusion Matrix (K=2)

	Sistem Informasi	Agroteknologi	ilmu hukum	ilmu komunikasi	ilmu pemerintahan	teknik arsitektur	teknik elektro	teknik informatika
Sistem Informasi	0	0	0	0	0	0	0	0
Agroteknologi	0	0	0	0	0	0	0	0
ilmu hokum	0	1	2	0	0	0	0	0
ilmu komunikasi	0	0	0	0	0	0	0	0
ilmu pemerintahan	0	0	0	1	0	0	0	0
teknik arsitektur	0	0	0	1	0	0	0	0
teknik elektro	1	0	1	0	0	0	0	0
teknik informatika	1	0	1	0	0	0	0	1

1. *Akurasi = 30 %*
2. *Precision = 30 %*
3. *Recall = 12,5%*

b. Uji Coba 2 (Nilai K =4)

Tabel 5.4 Hasil Uji Coba (K=4)

ID	Kelas Aktual	Kelas Prediksi
26	teknik arsitektur	teknik arsitektur
86	teknik informatika	teknik informatika
2	teknik elektro	teknik elektro
55	ilmu pemerintahan	ilmu hokum
75	ilmu hokum	Agroteknologi
93	ilmu hokum	ilmu hokum
16	ilmu hokum	ilmu hokum

73	teknik informatika	teknik informatika
54	teknik elektro	teknik informatika
95	teknik informatika	Sistem Informasi

Tabel 5.5 Confusion Matrix (K=4)

	Sistem Informasi	Agroteknologi	ilmu hukum	ilmu pemerintahan	teknik arsitektur	teknik elektro	teknik informatika
Sistem Informasi	0	0	0	0	0	0	0
Agroteknologi	0	0	0	0	0	0	0
ilmu hokum	0	1	2	0	0	0	0
ilmu pemerintahan	0	0	1	0	0	0	0
teknik arsitektur	0	0	0	0	1	0	0
teknik elektro	0	0	0	0	0	1	1
teknik informatika	1	0	0	0	0	0	2

1. *Akurasi = 60%*
2. *Precision = 60%*
3. *Recall = 40,47 %*

c. Uji Coba 2 (Nilai K =6)

Tabel 5.6 Hasil Uji Coba (K=6)

ID	Kelas Aktual	Kelas Prediksi
26	teknik arsitektur	teknik arsitektur
86	teknik informatika	teknik informatika
2	teknik elektro	teknik elektro
55	ilmu pemerintahan	ilmu komunikasi
75	ilmu hokum	manajemen
93	ilmu hokum	ilmu hukum
16	ilmu hokum	ilmu hukum
73	teknik informatika	teknik informatika
54	teknik elektro	teknik informatika
95	teknik informatika	teknik informatika

Tabel 5.7 Confusion Matrix (K=6)

	ilmu hukum	ilmu komunikasi	ilmu pemerintahan	manajemen	teknik arsitektur	teknik elektro	teknik informatika
ilmu hukum	2	0	0	1	0	0	0
ilmu komunikasi	0	0	0	0	0	0	0
ilmu pemerintahan	0	1	0	0	0	0	0
manajemen	0	0	0	0	0	0	0
teknik arsitektur	0	0	0	0	1	0	0
teknik elektro	0	0	0	0	0	1	1
teknik informatika	0	0	0	0	0	0	3

1. *Akurasi = 70%*
2. *Precision = 70%*
3. *Recall = 45,23 %*

d. Uji Coba 2 (Nilai K =8)

Tabel 5.8 Hasil Uji Coba (K=8)

ID	Kelas Aktual	Kelas Prediksi
26	teknik arsitektur	teknik arsitektur
86	teknik informatika	teknik informatika
2	teknik elektro	ilmu hukum
55	ilmu pemerintahan	teknik informatika
75	ilmu hokum	ilmu hukum
93	ilmu hokum	ilmu hukum
16	ilmu hokum	ilmu hukum
73	teknik informatika	teknik informatika
54	teknik elektro	teknik informatika
95	teknik informatika	teknik informatika

Tabel 5.9 Confusion Matrix (K=8)

	ilmu hukum	ilmu pemerintahan	teknik arsitektur	teknik elektro	teknik informatika
ilmu hokum	3	0	0	0	0
ilmu pemerintahan	0	0	0	0	1
teknik arsitektur	0	0	1	0	0
teknik elektro	1	0	0	0	1
teknik informatika	0	0	0	0	3

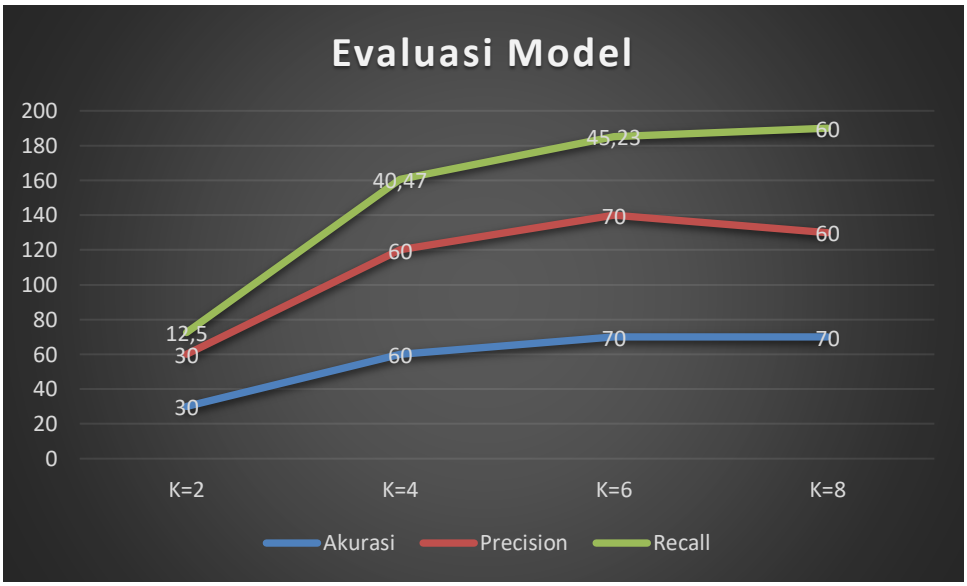
1. *Akurasi = 70%*
2. *Precision = 70%*
3. *Recall =60%*

Dari 4 kali uji coba yang telah di lakukan maka secara lengkap dapat dilihat pada tabel 5.10 Sebagai berikut :

Tabel 5.10 Hasil data uji

Uji Coba	Nilai K	Akurasi	Precision	Recall
1	2	30	30	12,5
2	4	60	60	40,47
3	6	70	70	45,23
4	8	70	70	60

Dari data tersebut tersebut di dapatkan akurasi dan precision tertinggi sebesar 70 untuk nilai K=6 dan 8, sedangkan untuk nilai recall tertinngi pada nilai K=8



Gambar 5.1 Evaluasi Model Terhadap Nilai K

5.2.5 Implementasi Model

Berikut adalah tampilan visualisasi yang di akses melalui terminal pada python untuk menginput data calon mahasiswa baru dengan beberapa variable sebagai berikut :

Gambar 5.10 Visualisasi Tampilan input data calon maba

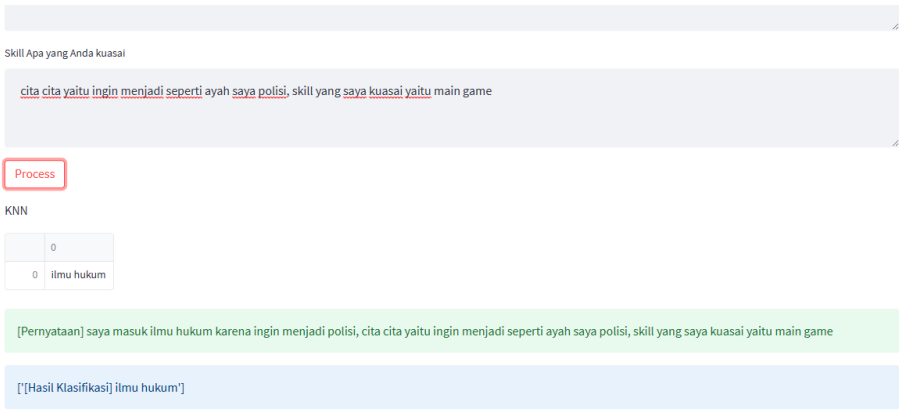
Pada tampilan di atas peneliti dapat menginput data calon mahasiswa baru tersebut dan berdasarkan pembobotan TF.IDF, dan juga perhitungan KNN serta klasifikasi dari Confusion Matrix, dapat diketahui calon mahasiswa tersebut lebih lebih cocok di jurusan apa. Berikut ini merupakan data calon mahasiswa baru yang akan diklasifikasi pada implementasi model.

Doc	Jawaban Peranyaan	Jurusan
1	saya ingi belajar ilmu informasi dan saya tertarik dengan jurusan sistem informasi,cita cita ingin menjadi programmer, skill yang punya yaitu memasak	?
2	saya masuk ilmu hukum karena ingin menjadi polisi, cita cita yaitu ingin menjadi seperti ayah saya polisi, skill yang saya kuasai yaitu main game	?

Jawaban calon mahasiswa diatas kemudian dimasukan kedalam text area satu-persatu untuk di klasifikasi agar bisa ditentukan jurusan yang termasuk pada kelas Teknik informatika,dkv,system informasi atau jurusan lainnya. Berikut ini merupakan gambar dari proses implementasi model pada streamlit untuk D1.

Berdasarkan gambar di atas maka dapat diketahui bahwa D1 termasuk pada Jurusan " Sistem Informasi"

Berikut ini merupakan gambar dari proses implementasi model pada streamlit untuk D2.



Berdasarkan gambar tersebut maka dapat diketahui bahwa D2 termasuk pada Jurusan "Ilmu HUkum".

BAB VI
KESIMPULAN DAN SARAN

6.1. KESIMPULAN

Dari hasi klasifikasi metode TF.IDF yang telah dilakukan maka dapat diketahui bahwa metode atau algoritma TF.IDF ini mampu membobotkan dengan baik sebuah kata dalam banyak dokumen, setelah dilakukan proses klasifikasi menggunakan algoritma *Term Frequency-Inverse Documet Frequency* dan K-Nearest Neighbor kita dapat melakukan kategori kelas atau jurusan mana yang paling cocok dengan calon mahasiswa tersebut karena memiliki tingkat akurasi dengan data yang diperoleh dari pelabelan sebanyak 100 data dan data uji sebanyak 10 data . Hal ini dilihat dari hasil evaluasi dengan *Confusion Matrikx*. Dengan beberapa kali percobaan, sehingga mendapatkan nilai K= 6 dann K=8 yang baik dengan akurasi yang dihasilkan yaitu sebesar 70%.

Jika peneliti menggunakan hanya Algoritma KNN saja tidak akan bisa, karena dalam penelitian ini data yang akan diolah hanya berupa teks, jadi untuk mendapatkan nilai dari masing-masing teks diperlukan TF.IDF untuk pembobotan kata sebagai fitur ekstrasi.

6.2. SARAN

dari penggunaan metode TF.IDF ini, jika tidak digunakan metode pembandingnya maka hasil akhirnya sangatlah sederhana, sehingga pada penelitian ini ditambahkan suatu metode seperti K-Nearest Neighbor sebagai pembanding untuk menyempurnakan penelitian ini. Selain itu metode KNN juga dapat menanggulangi jumlah data yang cukup besar dan metode ini juga mudah di implementasikan. tujuan digunakan metode K-Nearest Neighbor ialah untuk melihat tingkat kemiripan dari dokumen tes wawancara mahasiswa dari tiap jurusan dengan mahasiswa baru yang sudah memilih jurusan. Unntuk saran berikutnya bisa dipadukan dengan metode selain TF.IDF misalnya seperti *cosine similarity* atau metode lainnya.

DAFTAR PUSTAKA

[1] R. A. Sasmita, A. Z. Falani, F. I. Komputer, U. N. Surabaya, and T. Mining, "Pemanfaatan algoritma tf/idf pada sistem informasi ecomplaint handling," vol. 27, no. 1, pp. 27–33, 2018.

[2] M. Nurjannah and I. F. Astuti, "PENERAPAN ALGORITMA TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF) UNTUK TEXT MINING," vol. 8, no. 3, pp. 110–113, 2013.

[3] H. Zakiyudin and K. Marzuki, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF System Penerimaan Mahasiswa Baru pada Kampus Swasta Application of the Cosine Similarity Algorithm and Weighting of the TF-IDF System for New Student Admissions on Private Campuses," vol. 3, no. 1, pp. 19–27, doi: 10.30812/bite.v3i1.1110.

[4] U. Budiyanto and T. Fatimah, "Implementasi Algoritma Pembobotan TF-IDF dan Cosine Similarity untuk Penetapan Kategori Artikel pada Website Universitas Budi Luhur," vol. 10, pp. 218–223, 2022.

[5] N. Daulay, "Motivasi Dan Kemandirian Belajar Pada Mahasiswa Baru," *Al-Hikmah J. Agama dan Ilmu Pengetah.*, vol. 18, no. 1, pp. 21–35, 2021, doi: 10.25299/al-hikmah:jaip.2021.vol18(1).5011.

[6] H. Sari, G. L. Ginting, and T. Zebua, "Penerapan Algoritma Text Mining dan TF-IDF Untuk Pengelompokan Topik Skripsi Pada Aplikasi Repository STMIK Budi Darma," vol. 2, no. 7, pp. 414–432, 2021.

[7] S. Chamira, "Implementasi Metode Text Mining Frequency-Invers Document Frequency (Tf-Idf) Untuk Monitoring Diskusi Online," vol. 1, no. 3, pp. 97–102, 2022.

[8] A. Riyani, M. Zidny, and A. Burhanuddin, "Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen," vol. 2, no. 1, pp. 23–27, 2019.

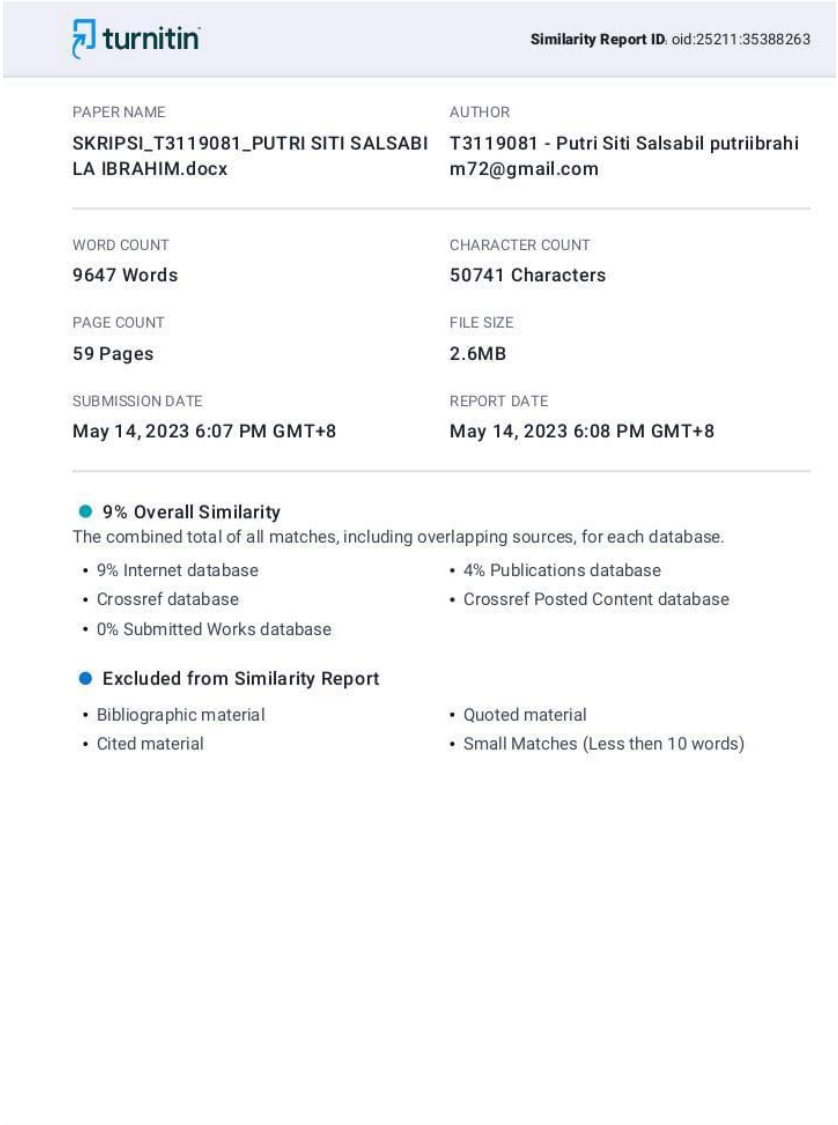
[9] Sesilia Novita R, Prihastuti Harsani, Arie Qur'ania" Penerapan K-Nearest Neighbor (KNN) untuk Klasifikasi Anggrek Berdasarkan Karakter Morfologi Daun dan Bunga" Vol.15,

No.1, Januari 2018, pp. 118−125

[10] Wiyli Yustanti " Algoritma K-Nearest Neighbour untuk Memprediksi Harga Jual Tanah" Vol. 9, No.1, 57-68, Juli 2012

Lampiran

Lampiran 1 :HASIL TURNITIN



turnitin

Similarity Report ID: oid:25211:35388263

PAPER NAME
SKRIPSI_T3119081_PUTRI SITI SALSABILA IBRAHIM.docx

AUTHOR
T3119081 - Putri Siti Salsabil putriibrahim72@gmail.com

WORD COUNT
9647 Words

CHARACTER COUNT
50741 Characters

PAGE COUNT
59 Pages

FILE SIZE
2.6MB

SUBMISSION DATE
May 14, 2023 6:07 PM GMT+8

REPORT DATE
May 14, 2023 6:08 PM GMT+8

● 9% Overall Similarity
The combined total of all matches, including overlapping sources, for each database.

- 9% Internet database
- 4% Publications database
- Crossref database
- Crossref Posted Content database
- 0% Submitted Works database

● Excluded from Similarity Report

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less then 10 words)

turnitin

Similarity Report ID: oid:25211:35388263

9% Overall Similarity

Top sources found in the following databases:

- 9% Internet database
- Crossref database
- 0% Submitted Works database
- 4% Publications database
- Crossref Posted Content database

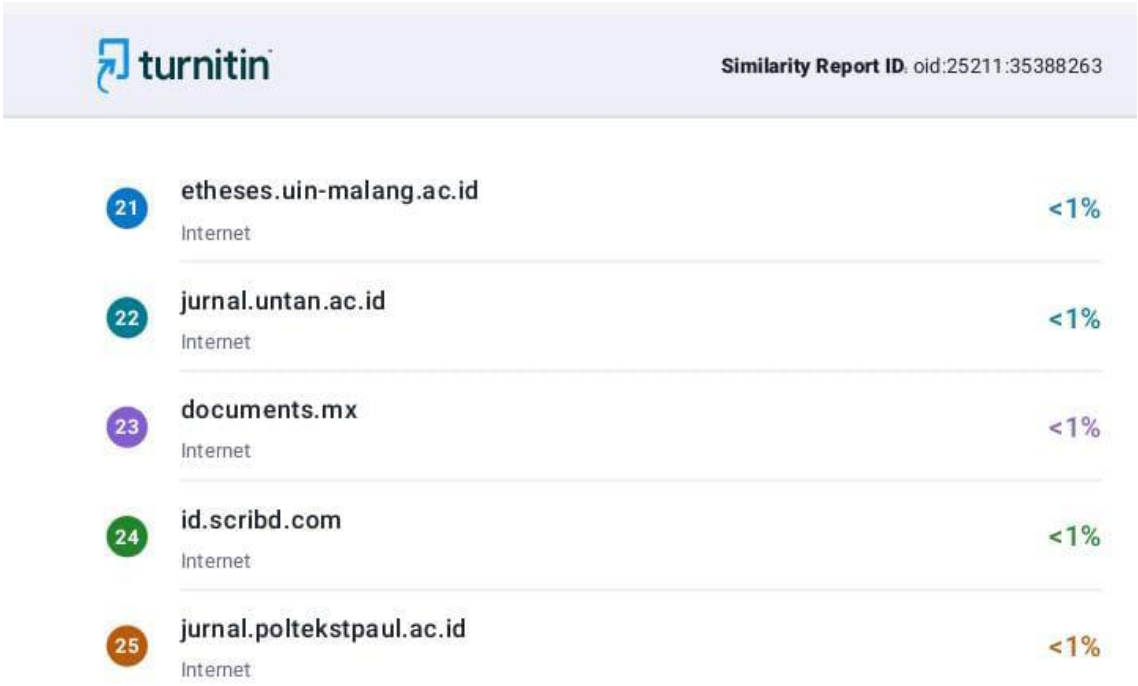
TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	journal.universitasbumigora.ac.id Internet	1%
2	jurnal-ticom.jakarta.aptikom.or.id Internet	<1%
3	researchgate.net Internet	<1%
4	dspace.uii.ac.id Internet	<1%
5	repository.uin-suska.ac.id Internet	<1%
6	wahl.lra-bgl.de Internet	<1%
7	journal.uinjkt.ac.id Internet	<1%
8	labdas.si.fti.unand.ac.id Internet	<1%

turnitin Similarity Report ID: oid:25211:35388263

9	tel.archives-ouvertes.fr Internet	<1%
10	123dok.com Internet	<1%
11	ejurnal.seminar-id.com Internet	<1%
12	ejurnal.unisan.ac.id Internet	<1%
13	inacl.id Internet	<1%
14	andi.ddns.net Internet	<1%
15	rstudio-pubs-static.s3.amazonaws.com Internet	<1%
16	docplayer.info Internet	<1%
17	download.garuda.kemdikbud.go.id Internet	<1%
18	smart.stmikplk.ac.id Internet	<1%
19	eprints.umm.ac.id Internet	<1%
20	text-id.123dok.com Internet	<1%



turnitin

Similarity Report ID: oid:25211:35388263

21	etheses.uin-malang.ac.id Internet	<1%
22	jurnal.untan.ac.id Internet	<1%
23	documents.mx Internet	<1%
24	id.scribd.com Internet	<1%
25	jurnal.poltekstpaul.ac.id Internet	<1%

Lampiran 2 : Listting Program

```
import operator
import time
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
from Sastrawi.StopWordRemover.StopWordRemoverFactory import
StopWordRemoverFactory
from sklearn.model_selection import train_test_split
import streamlit as st
from PIL import Image
import pandas as pd
import numpy as np
from sklearn.metrics import confusion_matrix
import matplotlib.pyplot as plt
import plotly.figure_factory as ff

from nltk.tokenize import RegexpTokenizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer(max_features=2500)


st.set_page_config(page_title='Pemilihan Jurusan', layout="wide")
image = Image.open('ichsan.JPEG')
st.image(image)
st.subheader('Pemilihan Jurusan')

excel_file = 'data.xlsx'
sheet_name = 'datamhs'

df = pd.read_excel(excel_file,
                   sheet_name=sheet_name,
                   header=0)
jumlahdata = df.No.count()
dataset = {}
dataset = {}
jurusan = {}
label = []
pertanyaan = []
for i in range(0, jumlahdata):
    dataset[df.No[i]] = df.Pertanyaan_1[i]
    jurusan[df.No[i]] = df.Jurusan_Sekarang[i]
    label.append(df.Jurusan_Sekarang[i])
    pertanyaan.append(df.Pertanyaan_1[i])
```

```
dnb = pd.DataFrame(columns=['pertanyaan', 'label'])
dnb['label'] = label
dnb['pertanyaan'] = pertanyaan
tab1, tab2, tab3 = st.tabs(["Tampilan Data", "Pembangunan Model",
"Implementasi Model"])

with tab1:
   st.header("Data Calon Mahasiswa")
   st.dataframe(df,width=100, height=1000,use_container_width=True)

with tab2:
   st.header("Pembangunan Model")
   if st.button('Mulai'):
            st.success("1. Preprocessing")
            # Case Folding
            for k in dataset.keys():

                hasil_case_folding = dataset[k].lower()
                dataset[k] = hasil_case_folding
            st.write("Case Folding:", dataset)

            # Tokenizing
            tokenizer = RegexpTokenizer(r"\w+")
            for k in dataset.keys():
                tokens = tokenizer.tokenize(dataset[k])
                dataset[k] = tokens
            st.write("Tokenisasi", dataset)

            # Number Removal
            for k in dataset.keys():
                tokens = []
                for t in dataset[k]:
                    if t.isnumeric() == False:
                        tokens.append(t)
                dataset[k] = tokens

            # Stopword Removal by Sastrawi
            factory = StopWordRemoverFactory()
            stopword_list = factory.get_stop_words()
            for k in dataset.keys():
                tokens = []
                for t in dataset[k]:
```

```
            if t not in stopword_list:
                tokens.append(t)
        dataset[k] = tokens
    st.write("Stopword Removing", dataset)

    # Stemming by Sastrawi
    factory = StemmerFactory()
    stemmer = factory.create_stemmer()
    for k in dataset.keys():
        tokens = []
        for t in dataset[k]:
            tokens.append(stemmer.stem(t))
        dataset[k] = tokens

    # Mengembalikan Format Dataset Awal
    st.write("stemming", dataset)
    for k in dataset.keys():
        dataset[k] = " ".join(dataset[k])
    st.success("2. Pembobotan Dokumen (TFIDF)")
    # Frekuensi Kemunculan Kata
    tf = CountVectorizer()
    term_doc_matrix = tf.fit_transform(dataset.values())
    pd.DataFrame(term_doc_matrix.toarray(), index=dataset.keys(
    ), columns=tf.get_feature_names_out())

    # Pembobotan TF-IDF
    vectorizer = TfidfVectorizer(max_features=2500)

    v_data = vectorizer.fit_transform(dataset.values()).toarray()
    st.write(v_data)


    st.success("4. KNN")
    X_train, X_test, y_train, y_test = train_test_split(
        v_data, dnb['label'], test_size=0.1, random_state=0)
    from sklearn.neighbors import KNeighborsClassifier
    knn = KNeighborsClassifier(n_neighbors=6)
    knn.fit(X_train, y_train)
    y_preds = knn.predict(X_test)
    st.info('Uji Coba Model')
    tes=pd.DataFrame(
         {
            "Data Aktual": y_test,
            "Data Prediksi": y_preds,
         }
```

```
            )
            tes
            #importing confusion matrix
            from sklearn.metrics import confusion_matrix
            confusion = confusion_matrix(y_test, y_preds)
            y_mrg=[*y_preds, *y_test]
            y_testcf = list(dict.fromkeys(y_mrg))
            sy_testcf=sorted(y_testcf)

            cm_df = pd.DataFrame(confusion,
                      index = [sy_testcf],
                      columns = [sy_testcf])
            st.write('Confusion Matrix\n')
            cm_df
            #importing accuracy_score, precision_score, recall_score,
f1_score
            from sklearn.metrics import accuracy_score, precision_score,
recall_score, f1_score
            st.info('\nAccuracy: {:.2f}\n'.format(accuracy_score(y_test,
y_preds)))


with tab3:
    pertanyaan1 = st.text_area('Mengapa memilih jurusan ini', )
    pertanyaan2 = st.text_area('Apa cita cita anda', )
    pertanyaan3 = st.text_area('Kelebihan apa yang ada punyai', )
    pertanyaan4 = st.text_area('Skill Apa yang Anda kuasai', )

    if st.button('Process'):
            # dataset["q"]=txt

            # Case Folding
            for k in dataset.keys():

                hasil_case_folding = dataset[k].lower()
                dataset[k] = hasil_case_folding
               # st.write("Case Folding:",dataset)
            # Tokenizing
            tokenizer = RegexpTokenizer(r"\w+")
            for k in dataset.keys():
                tokens = tokenizer.tokenize(dataset[k])
                dataset[k] = tokens
            # st.write("Tokenisasi",dataset)
            # Number Removal
            for k in dataset.keys():
```

```
            tokens = []
            for t in dataset[k]:
                if t.isnumeric() == False:
                    tokens.append(t)
            dataset[k] = tokens
        # Stopword Removal by Sastrawi
        factory = StopWordRemoverFactory()
        stopword_list = factory.get_stop_words()
        for k in dataset.keys():
            tokens = []
            for t in dataset[k]:
                if t not in stopword_list:
                    tokens.append(t)
            dataset[k] = tokens
        # st.write("Stopword Removing",dataset)
        # Stemming by Sastrawi
        factory = StemmerFactory()
        stemmer = factory.create_stemmer()
        for k in dataset.keys():
            tokens = []
            for t in dataset[k]:
                tokens.append(stemmer.stem(t))
            dataset[k] = tokens
        # Mengembalikan Format Dataset Awal
        # st.write("Stemming",dataset)
        for k in dataset.keys():
            dataset[k] = " ".join(dataset[k])
        # Frekuensi Kemunculan Kata
        tf = CountVectorizer()
        term_doc_matrix = tf.fit_transform(dataset.values())
        pd.DataFrame(term_doc_matrix.toarray(), index=dataset.keys(
        ), columns=tf.get_feature_names_out())

        # Pembobotan TF.IDF

        vectorizer = TfidfVectorizer(max_features=2500)
        v_data = vectorizer.fit_transform(dataset.values()).toarray()
        # st.write(v_data)

        X_train, X_test, y_train, y_test = train_test_split(
            v_data, dnb['label'], test_size=0.2, random_state=0)
        from sklearn.neighbors import KNeighborsClassifier
        knn = KNeighborsClassifier(n_neighbors=2)
        knn.fit(X_train, y_train)
        y_preds = knn.predict(X_test)
```

```
            tweet = pertanyaan1

            v_data = vectorizer.transform([tweet]).toarray()
            y_preds = knn.predict(v_data)



            st.write("KNN")
            from sklearn.neighbors import KNeighborsClassifier
            knn = KNeighborsClassifier(n_neighbors=4)
            knn.fit(X_train, y_train)
            y_preds = knn.predict(X_test)
            tweet = pertanyaan1

            v_data = vectorizer.transform([tweet]).toarray()
            y_preds = knn.predict(v_data)
            st.write(y_preds)
            # dengan asumsi bahwa 1 merupakan label positif
            st.success('[Pernyataan] '+pertanyaan1, )
            st.info('[Hasil  Klasifikasi] '+y_preds, )

    else:
        st.write('')
```

Lampiran 3 : Dataset

No	Nama_Mahasiswa	Jawaban Pertanyaan
1	Farhan Hamidjun	saya tertarik karena saya ingin melanjutkan jurusan dari sekolah multimedia makannya saya ambil jurusan dkv cita-cita saya ingin menjadi konsultan,saya suka foto,video,apapun yang berbaur seni, kelebihan saya road, saya bisa menggambar
2	Rahmawati putri m. razak	karena saya ingin belajar komputer,ingin menjadi programer,memiliki kelebihan tentang belajar komputer,bisa menggunakan komputer
3	Delpian dogomo	saya tertarik karena saya ingin jadi orang teknik dan rencana saya maau buka servis di papua,teknisi,rakit hardware,olahraga
4	anastasya z. lumoto	kemauan kedua orang tua,fotografer,editing,reporter,prisenter
5	Rinalwin poliama	yang membuat saya tertarik masuk jurusa teknik informatika adalah saya suka dengan komputer karena saya dari SMK sudah suka dengan komputer karena di SMK saya sudah dengan jurusan yang berkaitan dengan komputer,cita-cita saya adalah polisi,kelebihan saya adalah saya punya sill editing meskipun sedikit-sedikit,editing
6	Yolanda rahman	sesuai jurusan di sekolah ekonomi akuntansi ,menjadi auditing,Berhitung
7	sri devi husain	alasan saya memilih jurusan ini karena saya ingin lebih dalam untuk memperdalam jurusan ekonomi manajeme,pengusaha,pramuka
8	dwi agung blongkod	karena dengan keinginan saya sendiri karena nantinya setelah dari sini saya akan melanjutkan masuk polri,ingin menjadi anggota polri,yaitu sepak bola, olahraga,sepak bola
9	muhlis lamadi	karena saya ingin memperdalam tentang ilmu pertanian,cita-cita ingin memiliki kebun
10	rolan blongkod	karena ketertarikan dan mempunyai bakat,ingin jadi pelukis,menggambar,menggambar
11	Faris alamri	yang buat saya tertarik di manajemen saya mau menambah ilmu di bidang ini,pengusaha,kelebihan saya senang dalam ceramah
12	Maqbul sopansyah paputungan	prospek kedepannya sangat menjanjikan untuk lapangan pekerjaan,membahagiakan orang tua,dibidang game,buat game
13	rahmawaty adam	karena saya ingin mempelajari cara berkomunikasi dengan baik,prisenter,berkomunikasi dengan baik,bisa berinteraksi dengan baik
	Dst	....

Lampiran 4 :

Lampiran : RIWAYAT HIDUP PENELITI

Nama : Putri Siti Salsabila Ibrahim

Nim : T3119081

Tempat/Tanggal Lahir: Gorontalo, 22 Maret 2001

Email : putriibrahim72@gmail.com

Riwayat Pendidikan:

1. Tahun 2006, menyelesaikan Pendidikan TK di Taman Kanak-Kanak Indriahelbat
2. Tahun 2012, menyelesaikan Pendidikan di Sekolah Dasar Negeri 85 Kota Tengah, Kota Gorontalo
3. Tahun 2016, menyelesaikan Pendidikan di Sekolah Menengah Pertama Negeri 8 Gorontalo
4. Tahun 2019, menyelesaikan Pendidikan di Sekolah Menengah Kejuruan 1 Negeri Gorontalo
5. Tahun 2019, telah diterima menjadi Mahasiswa di Perguruan Tinggi Swasta Universitas Ichsan Gorontalo

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS ICHSAN GORONTALO
LEMBAGA PENELITIAN
Kampus Unisan Gorontalo Lt.3 - Jln. Achmad Nadjamuddin No. 17 Kota Gorontalo
Telp: (0435) 8724466, 829975 E-Mail: lembagapenelitian@unisan.ac.id

Nomor : 4417/PIP/LEMLIT-UNISAN/GTO/XI/2022
Lampiran : -
Hal : Permohonan Izin Penelitian

Kepada Yth,
Ketua Panitia Penerimaan Mahasiswa Baru Universitas Ichsan Gorontalo
di,-
Tempat

Yang bertanda tangan di bawah ini :

Nama	:	Dr. Rahmisyari, ST.,SE.,MM
NIDN	:	0929117202
Jabatan	:	Ketua Lembaga Penelitian

Meminta kesediannya untuk memberikan izin pengambilan data dalam rangka penyusunan **Proposal / Skripsi**, kepada :

Nama Mahasiswa	:	Putri Siti Salsabila Ibrahim
NIM	:	T3119081
Fakultas	:	Fakultas Ilmu Komputer
Program Studi	:	Teknik Informatika
Lokasi Penelitian	:	UNIVERSITAS ICHSAN GORONTALO
Judul Penelitian	:	PENERAPAN ALGORITMA TERM FREQUENCY - INVERSE DOCUMENT FREQUENCY UNTUK REKOMENDAASI PEMILIHAN JURUSAN MAHASISWA BARU (STUDI KASUS UNIVERSITAS ICHSAN GORONTALO)

Atas kebijakan dan kerja samanya diucapkan banyak terima kasih.

Gorontalo, 28 November 2022
Ketua,

Dr. Rahmisyari, ST.,SE.,MM
NIDN 0929117202

+

KEMENTERIAN RISET, TEKNOLOGI DAN PENDIDIKAN TINGGI
UNIVERSITAS ICHSAN GORONTALO
PANITIA PENERIMAAN MAHASISWA BARU T. A 2022-2023
Jln. Drs. Ahmad Nadjamuddin No 17 Tlp (0435) 829976 Gorontalo, Cp : 082292452423

SURAT KETERANGAN
NO : 077/UNISAN-G/VI/2021

Yang bertanda tangan dibawah ini :

Nama	: Sudirman Melangi, M.Kom
NIDN	: 0908017702
Jabatan	: Ketua Panitia Penerimaan Mahasiswa Baru Universitas Ichsan Gorontalo

Menyatakan,

Nama	: Putri Siti Salsabila Ibrahim
Nim	: T3119081
Prodi	: Teknik Informatika
Fakultas	: Ilmu Komputer

Bahwa yang bersangkutan diberikan izin penelitian di Universitas Ichsan Gorontalo dengan judul penelitian **"Penerapan Algoritma Term Frequency – Inverse Document Frequency Untuk Rekomendasi Pemilihan Jurusan Mahasiswa Baru**" waktu pelaksanaan penelitian dari Tanggal 01 Desember 2022 - 1 Februari 2023 Kepada caolon peneliti diharapkan:

1. Mematuhi segala ketentuan dan peraturan yang berlaku di Universitas Ichsan Gorontalo
2. Data penelitian bersifat rahasia dan tidak di izinkan menyampaikan informasi kepada pihak-pihak lain yang tidak berkepentingan.
3. Melaporkan kembali kepada Ketua Penerimaan Mahasiswa Baru Universitas Ichsan Gorontalo apabila pengambilan data telah berakhir.

Demikian surat keterangan ini dibuat untuk dapat dipergunakan sebagaimana mestinya,

Gorontalo, 03 Mei 2023

Ketua Panitia Penerimaan Mahasiswa Baru
Universitas Ichsan Gorontalo

Sudirman Melangi, M.Kom
NIDN. 09 080177 02

Tembusan:

1. Rektor Universitas Ichsan Gorontalo
2. Wakil Rektor I Bidang Akademik, Tata Kelola & Sistem Informasi
3. Mahasiswa yang bersangkutan
4. Arsip

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS ICHSAN GORONTALO
FAKULTAS ILMU KOMPUTER
SURAT KEPUTUSAN MENDIKNAS RI NOMOR 84/D/O/2001
Jl. Achmad Najamuddin No. 17 Telp. (0435) 829975 Fax (0435) 829976 Gorontalo

SURAT REKOMENDASI BEBAS PLAGIASI

No. 161/FIKOM-UIG/R/V/2023

Yang bertanda tangan di bawah ini :

Nama	: Irvan Abraham Salihi, M.Kom
NIDN	: 0928028101
Jabatan	: Dekan Fakultas Ilmu Komputer

Dengan ini menerangkan bahwa :

Nama Mahasiswa	:	Putri Siti Salsabila Ibrahim
NIM	:	T3119081
Program Studi	:	Teknik Informatika (S1)
Fakultas	:	Fakultas Ilmu Komputer
Judul Skripsi	:	Penerapan Algoritma Term Frequency Inverse Document Frequency Untuk Rekomendasi Pemilihan Jurusan Mahasiswa Baru

Sesuai hasil pengecekan tingkat kemiripan skripsi melalui aplikasi **Turnitin** untuk judul skripsi di atas diperoleh hasil *Similarity* sebesar **9%**, berdasarkan Peraturan Rektor No. 32 Tahun 2019 tentang Pendeteksian Plagiat pada Setiap Karya Ilmiah di Lingkungan Universitas Ichsan Gorontalo dan persyaratan pemberian surat rekomendasi verifikasi calon wisudawan dari LLDIKTI Wil. XVI, bahwa batas kemiripan skripsi maksimal 30%, untuk itu skripsi tersebut di atas dinyatakan **BEBAS PLAGIASI** dan layak untuk diujiankan.

Demikian surat rekomendasi ini dibuat untuk digunakan sebagaimana mestinya.

Gorontalo, 16 Mei 2023
Tim Verifikasi,

Zulfrianto Y. Lamasigi, M.Kom
NIDN. 0914089101

Mengetahui
Dekan,

Irvan Abraham Salihi, M.Kom
NIDN. 0928028101

Terlampir :
Hasil Pengecekan Turnitin

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS ICHSAN GORONTALO
FAKULTAS ILMU KOMPUTER
UPT. PERPUSTAKAAN FAKULTAS
SK. MENDIKNAS RI NO. 84/D/0/2001
Jl. Achmad Nadjamuddin No.17 Telp(0435) 829975 Fax. (0435) 829976 Gorontalo

SURAT KETERANGAN BEBAS PUSTAKA

No : 002/Perpustakaan-Fikom/V/2023

Perpustakaan Fakultas Ilmu Komputer (FIKOM) Universitas Ichsan Gorontalo dengan ini menerangkan bahwa :

Nama Anggota	:	Putri Siti Salsabila Ibrahim
No. Induk	:	T3119081
No. Anggota	:	M202330

Terhitung mulai hari, tanggal : Kamis, 04 Mei 2023, dinyatakan telah bebas pinjam buku dan koleksi perpustakaan lainnya.

Demikian keterangan ini di buat untuk di pergunakan sebagaimana mestinya.

Gorontalo, 04 Mei 2023
Mengetahui,
Kepala Perpustakaan

Apriyanto Alhamad, M.Kom
NIDN : 0924048601