MEMBANGUN CHATBOT UNTUK INFORMASI MAGANG DAN STUDI INDEPENDEN KAMPUS MERDEKA DENGAN ALGORITMA NAIVE BAYES

Cahya Diantoni, Ratna Mufidah, Heru Triana
Informatika, Universitas Singaperbangsa Karawang
Jl. HS. Ronggo Waluyo, Puseurjaya, Telukjambe Timur, Karawang, Jawa Barat, Indonesia
2010631170060@student.unsika.ac.id

ABSTRAK

Kampus Merdeka sebagai program lembaga pendidikan membutuhkan sistem informasi efisien untuk memenuhi kebutuhan mahasiswa terkait program Magang dan Studi Independen Bersertifikat (MSIB). *Respons* manual yang lambat dan potensi kesalahan manusiawi menjadi kendala utama dalam penyediaan informasi tepat waktu dan akurat. Oleh karena itu, penelitian ini berfokus pada pengembangan chatbot dengan algoritma Naive Bayes untuk mengatasi tantangan tersebut. Metode penelitian mengikuti pendekatan *Cross Industry Standard Process for Data Mining* (CRISP-DM), melibatkan pemahaman masalah, pengumpulan data, *preprocessing*, pembuatan model Naive Bayes, evaluasi model, dan implementasi di platform Telegram. Hasil penelitian menunjukkan model klasifikasi menggunakan algoritma Naive Bayes dengan tingkat akurasi mencapai 88,9%. Meskipun terdapat kendala data terbatas, algoritma ini dapat menangani distribusi frekuensi data yang tidak seimbang. Chatbot yang dikembangkan berpotensi meningkatkan akses mahasiswa, termasuk Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang (Fasilkom Unsika), terhadap informasi MSIB di Kampus Merdeka. Penelitian ini menegaskan bahwa chatbot dengan Naive Bayes dapat efektif menyediakan informasi secara cepat dan efisien. Algoritma ini terbukti efisien, terutama dalam mengatasi kendala data yang terbatas. Diharapkan hasil penelitian ini menjadi dasar untuk pengembangan lebih lanjut dalam implementasi kecerdasan buatan guna meningkatkan kualitas layanan informasi di lembaga pendidikan, termasuk Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang.

Kata kunci : *Algoritma Naïve Bayes, Chatbot, Fakultas Ilmu Komputer, Magang dan Studi Independen Bersertifikat, Sistem Informasi*

1. PENDAHULUAN

Dalam dunia pendidikan, layanan informasi yang dapat diakses langsung oleh mahasiswa tentu sangat dibutuhkan. Segala informasi terkait Kampus Merdeka seperti pendaftaran mahasiswa, program-program Kampus Merdeka, dan lain sebagainya dapat diakses dan tersedia di web resmi Kampus Merdeka (https://kampusmerdeka.kemdikbud.go.id/). Namun, untuk mendapatkan informasi lebih lengkap, pihak universitas hanya menyediakan kontak yang dapat dihubungi untuk memperjelas informasi dan menjawab pertanyaan-pertanyaan dari mahasiswa.

Dalam menjawab pertanyaan-pertanyaan mahasiswa secara manual, terdapat beberapa kendala yang dapat menghambat proses informasi terkait Kampus Merdeka ini. Pertama, waktu tunggu jawaban yang lama karena kesibukan staf akademik terlalu lama karena tergantung pada ketersediaan staf akademik yang sibuk dengan tugas lain. Kedua, adanya potensi kesalahan dalam memberikan informasi yang akurat dan konsisten. Dalam program Magang dan Studi Independen ini, detail-detail seperti syarat pendaftaran, jenis magang yang tersedia, atau batas waktu pendaftaran Studi Independen bisa sangat beragam. Staf akademik, meskipun berusaha sebaik mungkin, bisa saja memberikan informasi yang tidak sepenuhnya akurat. Ketiga, dalam kondisi tertentu, jumlah pertanyaan bisa menjadi sangat banyak sehingga terjadi penumpukan pertanyaan. Hal tersebut dapat menyebabkan ketidakpuasan mahasiswa terhadap layanan informasi yang lambat.

Layanan informasi Kampus Merdeka yang dapat diakses langsung oleh mahasiswa tentu akan memberikan dampak yang sangat positif bagi lembaga pendidikan, terutama dalam mengatasi kendala-kendala yang telah dijelaskan sebelumnya, seperti waktu respons yang lambat, potensi kesalahan manusiawi, dan lonjakan pertanyaan pada kondisi tertentu.

Salah satu contoh implementasi dari NLP adalah aplikasi chatbot [1]. Chatbot merupakan suatu program yang dapat melakukan percakapan melalui media tulisan atau pesan yang dapat merespon suatu perintah yang diberikan [2]. Tujuan utama dari Chatbot adalah untuk berinteraksi dengan manusia dan membantu dalam menyelesaikan berbagai tugas yang dapat mengurangi beban kerja manusia. Chatbot memiliki keunggulan karena dapat memberikan layanan 24/7 kepada pengguna, yang membuatnya menjadi alat yang sangat berguna dalam meningkatkan efisiensi layanan pelanggan untuk bisnis dan perusahaan [3]. Oleh karena itu, penggunaan Chatbot dapat menghasilkan peningkatan efisiensi yang signifikan.

Dalam penelitian Moechammad Sarosa dkk [4] menggunakan algoritma Naive Bayes dan Phase Reinforcement untuk menciptakan Chatbot yang dapat memberikan latihan wawancara dalam bahasa Inggris, terutama untuk wawancara kerja. Tujuannya adalah

memberikan solusi alternatif bagi pelajar yang kurang mahir dalam melakukan wawancara kerja dalam bahasa Inggris. Mereka menggunakan algoritma Naive Bayes untuk mengklasifikasikan hasil sesi wawancara antara pengguna dan interviewer-bot menjadi tiga kategori, yaitu permintaan, potensi, dan bakat. Klasifikasi kategori ini didasarkan pada perhitungan probabilitas dengan menggunakan Teorema Bayes. Hasil penelitian ini adalah pengembangan perangkat lunak dengan tingkat akurasi mencapai 86.93%.

Dalam penelitiannya Rena dkk [5] membangun aplikasi chatbot menggunakan algoritma naive bayes classifier untuk FAQ Grabads. Mereka telah menyiapkan dataset berisi pertanyaan umum (FAQ) sebanyak 10 pertanyaan dengan jawabannya. Dengan menggunakan split ratio sebesar 0,8 dan total 60 pertanyaan, mereka berhasil mencapai tingkat akurasi sebesar 93,33% dan tingkat kesalahan sebesar 6,66%.

Dari hasil studi literatur yang sudah dilakukan tersebut serta mengacu pada keberhasilan penelitian sebelumnya maka peneliti menjadikannya sebagai referensi dalam mengembangkan aplikasi Chatbot menggunakan algoritma Naive Bayes.

Berdasarkan uraian tersebut, tujuan dari penelitian ini adalah untuk mengembangkan chatbot yang efektif dalam menyediakan informasi terkait kampus merdeka, khususnya program Magang dan Studi Independen (MSIB) untuk mahasiswa Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang (Fasilkom Unsika). Dengan adanya chatbot yang dapat memberikan informasi yang akurat dan cepat, diharapkan mahasiswa Fasilkom Unsika akan lebih terbantu dalam mengambil keputusan terkait magang dan studi independen. Selain itu, penggunaan algoritma Naive Bayes sebagai bagian dari chatbot juga dapat meningkatkan kualitas interaksi dengan pengguna. Kerangka konseptual dalam penelitian ini melibatkan pengembangan chatbot menggunakan pendekatan algoritma Naive Bayes. Chatbot akan diarahkan untuk memahami pertanyaan pengguna melalui pemrosesan klasifikasi pertanyaan dan memberikan respons yang relevan berdasarkan informasi yang tersedia.

2. TINJAUAN PUSTAKA

2.1. Kampus Merdeka

Kampus merdeka merupakan lanjutan dari program merdeka belajar di tingkat pendidikan tinggi. Konsep ini bertujuan untuk mewujudkan SDM Unggul Indonesia dengan mengintegrasikan Profil Pelajar Pancasila. Perguruan tinggi diharapkan merancang pembelajaran inovatif melalui Kurikulum Merdeka Belajar Kampus (MBKM) agar mahasiswa mencapai pembelajaran optimal yang mencakup aspek kognitif, afektif, dan psikomotorik secara relevan [6].

2.1.1. Magang dan Studi Independen Bersertifikat

Program Magang dan Studi Independen Bersertifikat (MSIB) bertujuan memberikan kesempatan kepada mahasiswa untuk mengembangkan keterampilan, pengetahuan, dan sikap yang diperlukan di dunia industri melalui kombinasi bekerja dan belajar langsung pada proyek atau memecahkan masalah nyata [7].

2.2. Chatbot

Chatbot adalah program komputer yang dirancang untuk berkomunikasi dengan pengguna melalui media tulisan atau pesan. Dengan memanfaatkan kecerdasan buatan dan pemrosesan bahasa alami, chatbot dapat memahami perintah atau pertanyaan yang diberikan oleh pengguna dan memberikan respon yang sesuai [2].

2.3. Algoritma Naïve Bayes

Algoritma Naive Bayes merupakan suatu pendekatan klasifikasi dalam machine learning yang berasaskan pada teorema Bayes. Metode ini mengoperasikan dengan asumsi sederhana dan "naif," yaitu bahwa semua fitur atau kata yang terdapat dalam data input dianggap independen satu sama lain. Dengan kata lain, algoritma ini menganggap bahwa keberadaan suatu fitur tidak dipengaruhi oleh keberadaan fitur lainnya [8]. Hal ini membuat Algoritma Naive Bayes menjadi model yang efisien dan efektif, terutama dalam konteks analisis teks dan klasifikasi data.

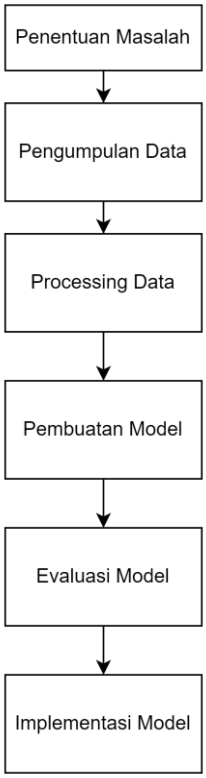
2.4. *Pre-Processing* Data

Pre-processing data mengacu pada langkah-langkah yang diambil untuk mempersiapkan dan membersihkan data sebelum diterapkan pada model atau algoritma tertentu [9].

3. METODE PENELITIAN

Penelitian ini menggunakan metodologi penelitian CRISP-DM. Metodologi CRISP-DM (Cross-Industry Standard Process for Data Mining) adalah sebuah metodologi yang digunakan dalam pengembangan solusi untuk masalah data mining [10].

Alasan menggunakan CRISP-DM dalam konteks ini adalah karena metodologi ini memberikan kerangka kerja yang sistematis dan terstruktur untuk mengelola proses penelitian data, dari pemahaman awal hingga implementasi. Hal ini dapat membantu memastikan bahwa hasil akhir proyek dapat berhasil diintegrasikan ke dalam platform chat seperti Telegram dengan efektif. Diagram Alur penelitian dapat dilihat pada Gambar 1.



Gambar 1. Metode penelitian CRISP-DM

Metodologi penelitian meliputi beberapa tahapan, yaitu:

3.1. Penentuan Masalah

Pada penelitian ini, masalah yang akan diselesaikan adalah bagaimana membangun chatbot informasi MSIB (Magang dan Studi Independen) yang dapat memberikan informasi tentang program MSIB Kampus Merdeka dan menjadi penghubung informasi dari panitia MBKM (Merdeka Belajar Kampus Merdeka) Fakultas Ilmu Komputer Unsika kepada mahasiswa. Chatbot ini diharapkan dapat membantu mahasiswa, calon peserta program MSIB Kampus Merdeka dalam mendapatkan informasi yang relevan dan cepat mengenai program MSIB Kampus Merdeka.

3.2. Pengumpulan data

Pada penelitian ini, data yang dikumpulkan adalah data informasi MSIB dan FAQ (Frequently Asked Questions) yang terkait dengan program MSIB. Data umum mengenai informasi MSIB dapat diakses dari website resmi program Kampus Merdeka dan data khusus yang berisi Informasi MSIB di Fasilkom Unsika didapatkan dari hasil wawancara dengan Panitia MBKM Fasilkom Unsika.

Data informasi MSIB yang dikumpulkan adalah data terbaru dan sedang berlaku sampai saat artikel ini ditulis. Data tersebut berisi 254 baris dataset dengan 2 atribut. Atribut yang terdapat dalam dataset adalah pertanyaan dan jawaban. Terdapat 35 nilai unik untuk jawaban, dan 254 nilai unik untuk pertanyaan. Adapun deskripsi dari atribut dataset yang digunakan dalam penelitian ini dapat dilihat pada Tabel 1.

Tabel 1. Deskripsi Dataset Penelitian

Atribut	Tipe	Deskripsi
Pertanyaan	*String*	Pertanyaan yang sering ditanyakan mengenai program MSIB Kampus Merdeka
Jawaban	*String*	Berisi informasi valid yang menjawab pertanyaan pertanyaan yang telah dikumpulkan

3.3. *Pre-processing* Data

Setelah data terkumpul, langkah berikutnya adalah melakukan *pre-processing* data untuk memastikan kebersihan dan kualitas data yang akan digunakan dalam pembuatan model. Teknik *pre-processing* data yang digunakan dalam penelitian ini adalah *tokenization*, *stopwords*, *lemmatization* dan *stemming* serta *bag of words*.

Berikut adalah penjelasan lengkap mengenai *pre-processing* data yang penulis gunakan dalam penelitian ini :

3.3.1. *Tokenization*

Tokenization adalah proses memecah teks menjadi unit-unit yang lebih kecil, yang disebut token [11] . Token bisa berupa kata, frasa, atau karakter, tergantung pada tingkat tokenisasi yang digunakan seperti yang terlihat pada Tabel 2.

Tabel 2. Contoh *Tokenization*

Teks	Hasil
“MSIB adalah singkatan dari Magang dan Studi Independen Bersertifikat.”	[“MSIB”, “adalah”, “singkatan”, “dari”, “Magang”, “dan”, “Studi”, “Independen”, “Bersertifikat”, “.”]

3.3.2. *Stopwords Removal*

Stopwords adalah kata-kata umum yang sering muncul dalam teks dan tidak memberikan informasi yang signifikan [12]. Teknik ini menghapus *stopwords* dari teks untuk meningkatkan kualitas analisis. Contoh kata *stopwords* : "dan", "dari", "atau". Proses *Stopwords Removal* dapat dilihat pada Tabel 3.

Tabel 3. Contoh *Stopwords Removal*

Teks	Hasil *Stopwords Removal*
“MSIB adalah singkatan dari Magang dan Studi Independen Bersertifikat.”	[“MSIB”, “adalah”, “singkatan”, “Magang”, “Studi”, “Independen”, “Bersertifikat”]

3.3.3. *Lemmatization* dan *Stemming*

Lemmatization dan *Stemming* adalah proses mengubah kata-kata dalam teks menjadi bentuk dasarnya (kata dasar) [13]. Misalnya, kata "mendapatkan" akan diubah menjadi "dapat". Hal tersebut dapat membantu dalam penggabungan variasi kata yang memiliki makna yang sama. Adapun contoh

dari *Lemmatization* dan *Stemming* dapat dilihat pada Tabel 4.

Tabel 4. Contoh *lemmatization* dan *stemming*

Teks	Hasil
"Mahasiswa mendapatkan peluang".	["Mahasiswa", "dapat", "peluang"]

3.3.4. *Bag of Words (BoW)*

Bag of Words adalah representasi teks dengan menghitung frekuensi kemunculan setiap kata dalam teks dan mengabaikan urutan kata [14]. Ini menghasilkan vektor berdasarkan kata-kata unik dalam teks terlihat dalam Tabel 5.

Tabel 5. Contoh *bag of words*

Contoh : terdapat 2 kalimat yaitu "Apa itu Magang?" dan "Apa itu Studi Independen"					
Kalimat	**Apa**	**itu**	**Magang**	**Studi**	**Independen**
Kalimat 1	1	1	1	0	0
Kalimat 2	1	1	0	1	1

3.4. Pembuatan model

Algoritma Naive Bayes adalah sebuah metode klasifikasi dalam machine learning yang didasarkan pada teorema Bayes dengan asumsi sederhana dan "naif" bahwa semua fitur (atau kata dalam teks) dalam data input adalah independen satu sama lain [8].

Pada tahap pembuatan model chatbot dengan algoritma Naive Bayes, model ini memanfaatkan probabilitas kemunculan kata dalam teks input untuk memprediksi kategori atau label yang paling mungkin [15]. Setiap kata dalam teks dihitung dalam konteks setiap label kategori, dan probabilitas tersebut digunakan untuk mengidentifikasi kategori dengan probabilitas tertinggi. Model Naive Bayes akan memilih label tersebut sebagai prediksi, sehingga chatbot akan merespons berdasarkan label yang paling sesuai dengan teks input pengguna.

3.5. Evaluasi model

Setelah berhasil membangun model chatbot informasi MSIB, langkah berikutnya adalah melakukan evaluasi untuk mengukur sejauh mana model tersebut dapat menyelesaikan masalah yang telah ditetapkan. Evaluasi dalam jurnal ini akan difokuskan pada metrik akurasi sebagai indikator utama kinerja chatbot.

Akurasi adalah metrik sederhana yang mengukur sejauh mana chatbot memberikan jawaban yang benar [16]. Rumusnya adalah:

$$\text{Akurasi} = \frac{Jawaban\ Benar}{Total\ Pertanyaan} \quad (1)$$

Dalam konteks evaluasi chatbot MSIB, akurasi memberikan gambaran langsung tentang seberapa efektif model dalam memberikan respons yang sesuai dengan pertanyaan pengguna. Dengan menekankan evaluasi akurasi, penelitian ini bertujuan untuk memberikan pemahaman yang jelas tentang kemampuan chatbot dalam memberikan jawaban yang benar dan sesuai dengan pertanyaan pengguna terkait informasi MSIB dan FAQ program MSIB. Evaluasi ini diharapkan dapat memberikan informasi yang cukup untuk menilai kinerja chatbot secara keseluruhan.

3.6. Implementasi model

Tahap implementasi chatbot dalam jurnal penelitian ini merupakan langkah penting dalam mewujudkan hasil penelitian ke dalam sebuah sistem yang praktis. Pada tahap ini, model chatbot yang telah dikembangkan diintegrasikan ke dalam server dengan memanfaatkan API Telegram. Implementasi dimulai dengan persiapan server yang mendukung bahasa pemrograman yang digunakan untuk membangun chatbot. Selanjutnya, semua dependensi yang diperlukan, termasuk library machine learning dan modul Telegram API, diinstal pada server. Setelah itu, sebuah bot Telegram dibuat dengan bantuan BotFather, yang memberikan token API yang diperlukan untuk menghubungkan chatbot dengan platform Telegram.

Langkah berikutnya adalah mengintegrasikan chatbot dengan Telegram API dengan menggunakan token yang diberikan oleh BotFather. Server juga dikonfigurasi agar dapat menerima permintaan dari Telegram API melalui pengaturan webhook atau endpoint yang sesuai. Kemudian, model chatbot yang telah dikembangkan diimplementasikan ke dalam server, termasuk pemuatan model, pengaturan logika chatbot, dan ketersediaan respons yang tepat.

Selanjutnya, chatbot diuji secara menyeluruh untuk memastikan kemampuan menerima perintah dari pengguna dan memberikan respons yang sesuai. Selama tahap ini, masalah dan kesalahan yang mungkin muncul selama pengujian diidentifikasi dan diperbaiki. Proses ini dapat melibatkan pengembangan lanjutan serta penyesuaian model chatbot jika diperlukan.

Terakhir, tahap implementasi ini melibatkan monitoring kinerja chatbot dan pemeliharaan berkala untuk memastikan bahwa bot tetap berjalan dengan baik. Seluruh proses diawali dengan pembuatan dokumentasi yang baik untuk pengguna akhir, dan jika diperlukan, publikasi aplikasi chatbot ini. Tahap implementasi chatbot ini menjadi jembatan penting yang menghubungkan penelitian dengan penerapan praktis, sehingga chatbot dapat memberikan nilai tambah yang nyata dalam konteks yang relevan.

Dalam keseluruhan tahapan tersebut, metodologi CRISP-DM sering digunakan sebagai panduan dalam melakukan penelitian machine learning.

4. HASIL DAN PEMBAHASAN

Pada penelitian ini menggunakan metodologi penelitian CRISP-DM karena hasil luaran penelitian ini dalam bentuk produk chatbot.

4.1. Penentuan Masalah

Dalam rangka penelitian ini, fokus utama adalah mengatasi tantangan dalam pembangunan chatbot informasi MSIB (Magang dan Studi Independen) yang berperan sebagai sumber informasi utama terkait program MSIB Kampus Merdeka. Tujuan utama adalah menciptakan sebuah chatbot yang efektif sebagai penghubung informasi antara panitia MBKM (Merdeka Belajar Kampus Merdeka) Fakultas Ilmu Komputer Unsika dengan mahasiswa. Diharapkan bahwa chatbot ini dapat memberikan bantuan yang optimal kepada mahasiswa dan calon peserta program MSIB Kampus Merdeka dengan menyediakan informasi yang relevan dan responsif mengenai berbagai aspek program tersebut.

4.2. Pengumpulan Data

Dalam proses pengumpulan data, peneliti menjalankan kolaborasi yang erat dengan panitia MBKM Fasilkom Unsika untuk merangkum dataset berupa pertanyaan dan jawaban yang relevan dengan program MSIB. Data umum mengenai MSIB diakses dari sumber resmi, yakni website program Kampus Merdeka. Sementara itu, informasi spesifik mengenai MSIB di Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang diperoleh melalui wawancara langsung dengan anggota panitia MBKM Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang.

Tabel informasi yang berhasil dikumpulkan terdiri dari 35 baris data dengan dua atribut utama, yaitu pertanyaan dan jawaban. Untuk memastikan relevansi dan akurasi data, penulis memastikan bahwa informasi yang diakses dan ditambahkan ke dalam dataset adalah yang paling terkini dan masih berlaku pada saat penulisan artikel ini.

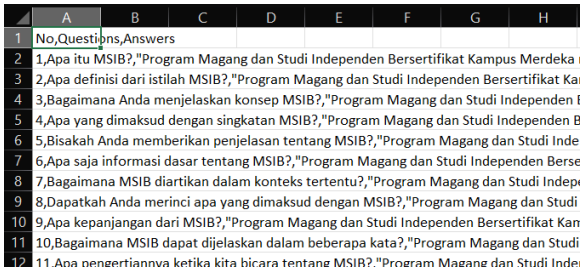
No	Questions	Answers
1	Apa itu Magang dan Studi Independen Bersertifikat?	Program Magang dan Studi Independen Bersertifikat Kampus Merdeka memberikan kesempatan kepada mahasiswa untuk mengasah dan mendapatkan kemampuan, pengetahuan dan sikap di dunia industri dengan cara bekerja dan belajar secara
2	Apakah program Magang atau Studi Independen diselenggarakan secara tatap muka di lokasi?	Program Magang atau Studi Independen dapat dilaksanakan secara luring, daring, atau hybrid. Hal tersebut menjadi pilihan masing-masing Mitra Industri.
3	Apa itu Magang?	Kegiatan dimana mahasiswa bekerja di organisasi mitra sebagai trainee selama

Gambar 2. Informasi MSIB

Kemudian dari dataset di atas penulis memperluas data untuk pertanyaan, dengan cara membuat pertanyaan yang mirip dan sesuai untuk beberapa jawaban yang sama. Dari hasil memperbanyak pertanyaan untuk 35 Jawaban unik, didapatkan dataset baru dengan jumlah 1164 baris.

Dataset ini dibagi menjadi dua jenis dataset, yaitu untuk pelatihan model dan validasi. Penulis membaginya menjadi 910 dataset untuk tahap pelatihan model dan 254 dataset untuk tahap validasi, kemudian dataset tersebut dikonversi kedalam format CSV. Proses ini bertujuan untuk memastikan bahwa model yang dibangun memiliki keberlanjutan dan daya prediksi yang baik terhadap pertanyaan yang mungkin diajukan oleh mahasiswa terkait MSIB.

Langkah-langkah ini menjamin bahwa dataset yang dihasilkan tidak hanya akurat dan up-to-date, tetapi juga mencerminkan kebutuhan mahasiswa secara real-time. Kolaborasi erat dengan panitia MBKM dan referensi dari sumber resmi menambah kredibilitas dan validitas dataset yang menjadi dasar pembangunan chatbot untuk menyediakan informasi MSIB secara efektif kepada mahasiswa.



Gambar 3. Dataset dengan format CSV

Gambar 4 mencakup kumpulan data pertanyaan yang sering diajukan (FAQ) yang telah diolah sehingga dapat dioperasikan dengan menggunakan algoritma *Naive Bayes Classifier*. Data ini telah diolah dan diorganisir dalam format CSV (*Comma-Separated Values*), memudahkan proses analisis dan implementasi algoritma klasifikasi. Dengan struktur CSV, setiap pertanyaan dan informasi terkaitnya dapat dengan mudah diakses dan digunakan oleh sistem untuk melatih dan menguji model klasifikasi Naive Bayes. Format CSV juga memungkinkan interoperabilitas yang baik dengan berbagai platform dan alat analisis data.

4.3. *Pre-processing Data*

Setelah berhasil mengumpulkan data, langkah selanjutnya adalah menjalankan proses pre-processing data guna memastikan kebersihan dan kualitas data yang akan menjadi dasar dalam pembuatan model. Penelitian ini menerapkan beberapa teknik pre-processing data, antara lain tokenization, removal stopwords, lemmatization dan stemming, serta pendekatan bag of words.

Sebelum dapat menerapkan algoritma Naive Bayes Classifier pada dataset, prapemrosesan (pre-processing) diperlukan untuk mengoptimalkan kualitas data. Prapemrosesan ini melibatkan dua tahap utama pada kalimat-kalimat pertanyaan dalam dataset, yaitu tokenisasi dan penghapusan stopwords.

Pada langkah tokenisasi, terdapat dua proses yang harus dilakukan. Pertama, semua huruf besar pada teks akan diubah menjadi huruf kecil (lowercase). Kedua, teks akan dibagi menjadi kumpulan kata tanpa memperhatikan keterhubungan antar kata satu dengan yang lain, serta peran dan posisinya dalam kalimat.

Karakter diterima dalam kumpulan kata menurut urutan abjad.

Sementara pada langkah penghapusan stopwords, jika sebuah kata dalam dataset terdapat pada daftar stopwords, maka kata tersebut akan dihilangkan. Namun, jika kata tersebut tidak terdapat dalam daftar stopwords, proses akan berlanjut tanpa menghilangkan kata tersebut.

```
# Fungsi untuk melakukan tokenisasi dan penghapusan stopwords
def preprocess_text(text):
    # Tokenisasi
    tokens = word_tokenize(text.lower())

    # Penghapusan stopwords
    stop_words = set(stopwords.words('indonesian'))
    filtered_tokens = [word for word in tokens if word.isalnum() and word not in stop_words]

    return filtered_tokens
```

Gambar 4. Code proses stopwords

Penggunaan tokenisasi dan penghapusan stopwords dalam pra-pemrosesan teks membantu menyederhanakan representasi teks, sehingga *Naive Bayes Classifier* dapat lebih efektif memahami dan memproses informasi relevan yang diperlukan untuk klasifikasi pertanyaan pada dataset. Kemudian dari kedua proses tersebut didapatkan contoh data.

No	Pertanyaan Sebelum	Pertanyaan Sesudah
1	Apa itu MSIB?	msib
2	Apa definisi dari istilah MSIB?	definisi istilah msib
3	Bagaimana Anda menjelaskan konsep MSIB?	konsep msib
4	Apa yang dimaksud dengan singkatan MSIB?	singkatan msib
5	Bisakah Anda memberikan penjelasan tentang MSIB?	penjelasan msib
6	Apa saja informasi dasar tentang MSIB?	informasi dasar msib
7	Bagaimana MSIB diartikan dalam konteks tertentu?	msib diartikan konteks
8	Dapatkah Anda merinci apa yang dimaksud dengan MSIB?	dapatkah merinci msib
9	Apa kepanjangan dari MSIB?	kepanjangan msib
10	Bagaimana MSIB dapat dijelaskan dalam beberapa kata?	msib

Gambar 5. Hasil proses stopwords

Gambar 5 merupakan contoh data dalam bentuk kalimat setelah dilakukan proses tokenization dan stopwords. Contoh kalimatnya ”Apa itu MSIB?” menjadi ”MSIB”

Sebelumnya, dataset melalui tahap tokenisasi dan penghapusan stopwords. Sekarang, setiap kata dalam kalimat akan diberikan tanda (tagging) untuk mengekstrak kata kerja dan kata benda. Proses selanjutnya melibatkan lemmatisasi dan stemming, di mana kata-kata dalam kalimat akan diubah menjadi bentuk dasar mereka dan kemudian dikelompokkan.

Pada metode stemming, terjadi pemisahan imbuhan seperti awalan dan akhiran untuk mendapatkan kata dasar. Untuk meningkatkan efektivitas, daftar kata yang akan dihilangkan telah disusun dan disimpan dalam ruang stem list. Selanjutnya, data akan dibandingkan dengan stem list. Jika ada kesamaan, data tersebut dianggap bising dan dihapus.

Proses ini bertujuan untuk menyederhanakan dan membersihkan kalimat agar dapat memahami kata-kata asalnya (melalui stemming) dan mengelompokkan kata-kata menjadi satu item (melalui lemmatisasi). Penggunaan stem list membantu mengidentifikasi kata-kata yang dianggap tidak relevan atau berisik, sehingga dapat dihilangkan dari dataset.

```
# Stemming (menggunakan Sastrawi) dan Lemmatisasi (WordNetLemmatizer)
processed_tokens = [lemmatizer.lemmatize(stemmer.stem(word)) for word in words]

# Gabungkan kembali kata-kata yang telah diolah
preprocessed_text = ' '.join(processed_tokens)
```

Gambar 6. Code proses stemming dan lemmatisasi

Lematisasi dan stemming adalah teknik pra-pemrosesan teks yang bertujuan untuk menyederhanakan kata-kata dalam sebuah dokumen. Stemming menghilangkan awalan atau akhiran kata untuk menghasilkan bentuk dasar atau kata dasar, sementara lemmatisasi mengubah kata-kata ke bentuk dasar berdasarkan kamus atau aturan linguistik yang memahami makna kata. Meskipun stemming lebih cepat dan sederhana, lemmatisasi cenderung memberikan hasil yang lebih akurat karena melibatkan analisis linguistik yang lebih mendalam. Pilihan antara keduanya tergantung pada kebutuhan tugas pemrosesan teks, dengan stemming lebih cocok untuk situasi di mana kecepatan dan sederhana diperlukan, sedangkan lemmatisasi lebih cocok untuk tugas yang memerlukan pemahaman kata-kata yang lebih mendalam dan akurat.

No	Pertanyaan Sebelum	Pertanyaan Setelah (Stemming dan Lemmatisasi)
1	Apa itu MSIB?	msib
2	Apa definisi dari istilah MSIB?	definisi istilah msib
3	Bagaimana Anda menjelaskan konsep MSIB?	konsep msib
4	Apa yang dimaksud dengan singkatan MSIB?	singkat msib
5	Bisakah Anda memberikan penjelasan tentang MSIB?	jelas msib
6	Apa saja informasi dasar tentang MSIB?	informasi dasar msib
7	Bagaimana MSIB diartikan dalam konteks tertentu?	msib arti konteks
8	Dapatkah Anda merinci apa yang dimaksud dengan MSIB?	dapat merinci msib
9	Apa kepanjangan dari MSIB?	panjang msib
10	Bagaimana MSIB dapat dijelaskan dalam beberapa kata?	msib

Gambar 7. Hasil proses stemming dan lemmatisasi

Gambar 7 memperlihatkan contoh data yang telah melalui tahap lemmatisasi dan stemming, di mana proses ini melibatkan ekstraksi fitur dari setiap kalimat. Selanjutnya, data tersebut akan menjalani proses pembentukan model Bag of Words. Dengan pendekatan yang sederhana, Bag of Words merepresentasikan sebuah kalimat sebagai sekumpulan kata-kata, mengabaikan urutan spesifik kemunculan kata-kata tersebut. Hasil ini menciptakan representasi numerik yang dapat digunakan dalam analisis dan klasifikasi teks, memudahkan pemahaman dan pengolahan informasi tanpa mempertimbangkan struktur urutan kata dalam kalimat. Dalam konteks jurnal, langkah ini mendukung analisis fitur teks yang berguna dalam memahami esensi dan makna dari konten yang diolah, membantu pemahaman informasi yang relevan dalam dataset teks yang kompleks.

4.4. Pembuatan Model

Algoritma Naive Bayes, sebuah metode klasifikasi dalam machine learning, mendasarkan prediksinya pada teorema Bayes dengan asumsi sederhana dan "naif" bahwa semua fitur (atau kata dalam teks) dalam data input adalah independen satu sama lain [5]. Pada tahap pembuatan model chatbot menggunakan algoritma Naive Bayes, model ini memanfaatkan probabilitas kemunculan kata dalam teks input untuk memprediksi kategori atau label yang paling mungkin. Setiap kata dalam teks dihitung dalam konteks setiap label kategori, dan probabilitas

tersebut digunakan untuk mengidentifikasi kategori dengan probabilitas tertinggi. Model Naive Bayes akan memilih label tersebut sebagai prediksi, sehingga chatbot dapat merespons berdasarkan label yang paling sesuai dengan teks input pengguna.

Dalam pengembangan lebih lanjut, implementasi algoritma Naive Bayes pada tahap ini dapat dilihat pada kode di bawah ini:

```
# Load dataset
df = pd.read_csv('training_info.csv')

# Preprocess dataset
tfidf_vectorizer = TfidfVectorizer()
X = tfidf_vectorizer.fit_transform(df['Questions'])
y = df['Answers']

# Train a classifier
classifier = MultinomialNB()
classifier.fit(X, y)
```

Gambar 8. Code training model

Pada kode di atas, dataset training diunduh dan diproses menggunakan TfidfVectorizer untuk mengubah teks menjadi representasi vektor dengan bobot term frequency-inverse document frequency (TF-IDF). Selanjutnya, model klasifikasi Naive Bayes MultinomialNB dilatih menggunakan fitur dan label yang telah dihasilkan.

```
# Function to get an answer
def get_answer(user_input):
    user_input_tfidf = tfidf_vectorizer.transform([user_input])
    answer = classifier.predict(user_input_tfidf)[0]
    return answer

# Test the chatbot
user_input = "bagaimana cara mendapatkan sptjm?"
answer = get_answer(user_input)
print("Jawaban:", answer)
```

Gambar 9. Code untuk mendapatkan jawaban

Fungsi `get_answer` digunakan untuk memprediksi jawaban berdasarkan input pengguna, dan hasilnya diuji dengan pertanyaan "bagaimana cara mendapatkan sptjm?". Jawaban yang diprediksi kemudian ditampilkan dalam output program.

```
Jawaban: Tentang SPTJM
1. Unduh Template (http://ringkas.kemdikbud.go.id/SPTJMMSIB5)
2. Isi dengan valid, Isi data diri, kondisi menerima beasiswa lain atau tidak, n
3. SPTJM ditanda tangani Dekan Fakultas,
Dr. Mayasari, S.S., M.Hum
NIP. 197909262021212005
4. Unggah SPTJM, unggah scan SPTJM (dengan kondisi yang baik dan terang), sudah
5. Folder SPTJM, unggah versi .docx ke folder https://bit.ly/uploadSPTJM_MSIB5
6. Unduh SPTJM, ada di folder https://bit.ly/unduhSPTJM_MSIB
```

Gambar 10. Hasil jawaban dari chatbot

4.5. Evaluasi Model

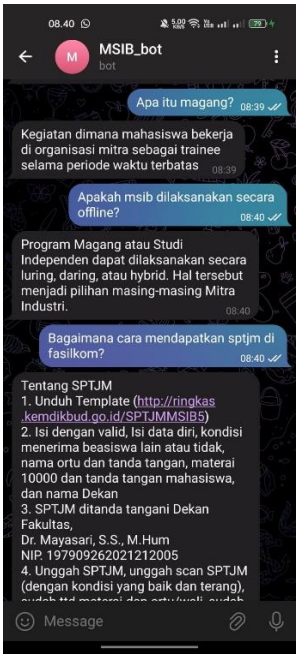
Dalam penelitian ini, penulis telah melakukan evaluasi kinerja chatbot dengan menggunakan dataset evaluasi yang terdiri dari 254 pertanyaan yang telah diberi label jawaban yang benar. Hasil evaluasi tersebut menunjukkan bahwa chatbot berhasil memberikan jawaban yang benar sebanyak 226 pertanyaan, atau setara dengan tingkat akurasi sebesar 88,9%. Akurasi ini mencerminkan sejauh mana model chatbot dapat menanggapi pertanyaan pengguna dengan respons yang tepat dan sesuai dengan harapan. Hasil evaluasi ini memberikan gambaran positif tentang kemampuan chatbot dalam memberikan jawaban yang akurat dalam konteks informasi MSIB dan FAQ program MSIB. Evaluasi ini dapat dijadikan dasar untuk menilai efektivitas dan kualitas respons chatbot dalam memenuhi kebutuhan pengguna terkait informasi yang dicari.

```
# Menghitung jumlah data yang benar
jumlah_benar = (y_validasi == prediksi_validasi).sum()

# Menghitung total jumlah data
total_data = len(y_validasi)

# Menampilkan jumlah data yang benar dan total jumlah data
print("Jumlah Data yang Benar:", jumlah_benar)
print("Total Data:", total_data)

Jumlah Data yang Benar: 226
Total Data: 254
```

Gambar 11. Pengujian hasil jawaban chatbot

4.6. Implementasi Model

Tahap implementasi model chatbot menjadi poin krusial dalam menghadirkan hasil penelitian ke dalam lingkungan praktis. Hasil dari implementasi chatbot dapat diakses melalui aplikasi Telegram dengan url *https://t.me/MSIB_Chatbot* Berikut adalah hasil implementasi chatbot melalui aplikasi Telegram, pada gambar X .



Gambar 12. Hasil deployment chatbot pada telegram

Tahap implementasi ini membentuk jembatan penting yang menghubungkan hasil penelitian dengan penerapan praktis, memungkinkan chatbot memberikan nilai tambah yang nyata dalam konteks yang relevan.

5. KESIMPULAN DAN SARAN

Kesimpulan dari penelitian ini adalah berhasilnya pengembangan chatbot menggunakan algoritma Naive Bayes untuk menyediakan informasi terkait program Magang dan Studi Independen di Kampus Merdeka, khususnya di Fakultas Ilmu Komputer Unsika. Metodologi CRISP-DM digunakan secara efektif dalam tahapan pengembangan, mulai dari penentuan masalah hingga implementasi. Evaluasi kinerja model menunjukkan tingkat akurasi yang baik sebesar 88,9%. Tahap implementasi chatbot berhasil dilakukan melalui Telegram, memberikan kontribusi positif terhadap pengalaman pengguna dalam mendapatkan informasi terkait MSIB. Meskipun terdapat batasan terkait jumlah data, algoritma Naive Bayes terbukti efisien, dan dengan peningkatan jumlah data serta tuning model, akurasi chatbot dapat ditingkatkan lebih lanjut. Chatbot ini memiliki potensi besar untuk meningkatkan kualitas interaksi dengan pengguna dan memberikan kontribusi positif dalam penyediaan informasi di lembaga pendidikan. Sebagai saran, penelitian ini dapat menjadi dasar untuk pengembangan lebih lanjut dalam pemanfaatan teknologi AI guna meningkatkan layanan informasi di lingkungan pendidikan.

DAFTAR PUSTAKA

[1] N. A. Purwitasari and M. Soleh, "Implementasi Algoritma Artificial Neural Network Dalam Pembuatan Chatbot Menggunakan Pendekatan Natural Language Parocessing," *Jurnal IPTEK*, vol. 6, no. 1, pp. 14–21, 2022, doi: 10.31543/jii.v6i1.192.

[2] A. Hikmah, F. Azmi, and R. A. Nugrahaeni, "Implementasi Natural Language Processing Pada Chatbot Untuk Layanan Akademik," *e-Proceeding of Engineering*, vol. 10, no. 1, pp. 371–382, 2023.

[3] M. I. P. N. Mayla Humaira As-syiva, "Analisis Peran Chatbot dalam Meningkatkan Pelayanan Terhadap Konsumen di E-Commerce," *Jurnal Multidisiplin Saintek*, vol. 1, no. 0, 2023.

[4] M. Sarosa, M. Kusumawardani, A. Suyono, and Z. Sari, "Implementasi Chatbot Pembelajaran Bahasa Inggris menggunakan Media Sosial," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 6, no. 3, p. 317, 2020, doi: 10.26418/jp.v6i3.43191.

[5] R. C. Hutama, F. Fauziah, and R. T. Komalasari, "Aplikasi Chatbot Berbasis Teks Menggunakan Algoritma Naive Bayes Classifier FAQ GrabAds," *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, vol. 6, no. 1, p. 90, 2021, doi: 10.30998/string.v6i1.9919.

[6] R. Vhalery, A. M. Setyastanto, and A. W. Leksono, "Kurikulum Merdeka Belajar Kampus Merdeka: Sebuah Kajian Literatur," *Research and Development Journal of Education*, vol. 8, no. 1, p. 185, 2022, doi: 10.30998/rdje.v8i1.11718.

[7] A. Rahman, D. C. Sukmajati, M. Mawar, E. Satispi, and D. Gunanto, "Implementasi Kebijakan pada Program Magang dan Studi Independen Bersertifikat di Indonesia," *SOSIOHUMANIORA: Jurnal Ilmiah Ilmu Sosial Dan Humaniora*, vol. 9, no. 2, pp. 266–291, 2023, doi: 10.30738/sosio.v9i2.14832.

[8] A. T. Gurinder Singh, Bhawna Kumar, Loveleen Gaur, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," *2019 International Conference on Automation, Computational and Technology Management (ICACTM), London, UK*, pp. 593–596, 2019, doi: 10.1109/ICACTM.2019.8776800.

[9] S. Saifullah, M. Zarlis, Z. Zakaria, and R. W. Sembiring, "Analisa Terhadap Perbandingan Algoritma Decision Tree Dengan Algoritma Random Tree Untuk Pre-Processing Data," *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, vol. 1, no. 2, p. 180, 2017, doi: 10.30645/j-sakti.v1i2.41.

[10] Y. A. Singgalen, "… Metode CRISP-DM dalam Klasifikasi Data Ulasan Pengunjung Destinasi Danau Toba Menggunakan Algoritma Naïve Bayes Classifier (NBC) dan Decision Tree (DT)," *Jurnal Media Informatika Budidarma*, vol. 7, pp. 1551–1562, 2023, doi: 10.30865/mib.v7i3.6461.

[11] G. N. R Prasad Sr Asst professor, "Identification of Bloom's Taxonomy level for the given Question paper using NLP Tokenization technique," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 13, pp. 1872–1875, 2021.

[12] K. S. K. V. Ghag, "Comparative analysis of effect of stopwords removal on sentiment classification," *IEEE International Conference on Computer Communication and Control (IC4-2015)*, 2015, [Online]. Available: https://doi.org/10.1109/IC4.2015.7375527

[13] A. W. Pradana and M. Hayaty, "The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 4, pp. 375–380, 2019, doi: 10.22219/kinetik.v4i4.912.

[14] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," *PLoS One*, vol. 15, no. 5, pp. 1–22, 2020, doi: 10.1371/journal.pone.0232525.

[15] R. Al Habsi, R. A. D. Anggoro, M. A. Valio, Y. Widiastiwi, and N. Chamidah, "Analisis Sentimen Terhadap Vaksin Covid-19 Di Jejaring Sosial Twitter Menggunakan Algoritma Naïve Bayes," *Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer dan Aplikasinya*, vol. 2, no. 2, pp. 239–248, 2021, [Online]. Available:

https://conference.upnvj.ac.id/index.php/senamika/article/view/1714

[16] B. Falah and Nerisma Eka Putri, "Artificial Intelligence Berbasis Chatbot: Sarana Baru Panduan Hukum Keluarga Digital," *QISTHOSIA : Jurnal Syariah dan Hukum*, vol. 4, no. 2, pp. 126–140, 2023, doi: 10.46870/jhki.v4i2.765.