

# Implementasi Metode TF-IDF dan *Cosine Similarity* pada *Chatbot* Layanan Informasi Madrasah

<sup>1\*</sup>Rina Padila Febriani <sup>2</sup>Ahmad Bagus Setiawan

<sup>1,2</sup> Teknik Informatika, Universitas Nusantara PGRI Kediri

E-mail: <sup>1</sup>[rinapadila27@gmail.com](mailto:rinapadila27@gmail.com) <sup>2</sup>[ahmad.bagus@unpkediri.ac.id](mailto:ahmad.bagus@unpkediri.ac.id)

**Penulis Korespondens : Rina Padila Febriani**

**Abstrak**— Perkembangan teknologi informasi mendorong lembaga pendidikan untuk meningkatkan layanan informasi secara digital. Penelitian ini bertujuan merancang dan membangun *Chatbot* layanan informasi pendaftaran siswa baru berbasis *Natural Language Processing* pada MA Bahrul Ulum Tanjungjaya Subang menggunakan metode TF-IDF dan *Cosine similarity*. Dataset yang digunakan berupa *Frequently Asked Questions* yang berkaitan dengan layanan pendaftaran madrasah. Sistem dikembangkan menggunakan bahasa pemrograman Python dan framework Flask. Pengujian dilakukan menggunakan 80 pertanyaan dari delapan kategori layanan informasi. Hasil pengujian menunjukkan *Chatbot* mampu memberikan jawaban yang sesuai pada 63 pertanyaan dengan tingkat keberhasilan sebesar 78,75 persen. Hasil tersebut menunjukkan bahwa metode TF-IDF dan *Cosine similarity* cukup efektif digunakan pada *Chatbot* layanan informasi berbasis web.

**Kata Kunci**— *Chatbot*, *Cosine similarity*, natural language processing, pendaftaran siswa baru, TF-IDF

**Abstract**— The development of information technology encourages educational institutions to improve digital information services. This study aims to design and develop a *Chatbot* for new student admission information services based on *Natural Language Processing* at MA Bahrul Ulum Tanjungjaya Subang using TF-IDF and *Cosine similarity* methods. The dataset used consists of *Frequently Asked Questions* related to school admission services. The system was developed using Python programming language and Flask framework. System testing was conducted using 80 questions from eight categories of information services. The results showed that the *Chatbot* successfully provided appropriate answers to 63 questions with an accuracy rate of 78.75 percent. These results indicate that TF-IDF and *Cosine similarity* are effective methods for web-based information service *Chatbots*.

**Keywords**— *Chatbot*, *Cosine similarity*, natural language processing, new student admission, TF-IDF

This is an open access article under the CC BY-SA License.



## I. PENDAHULUAN

Perkembangan teknologi informasi mendorong lembaga pendidikan untuk meningkatkan kualitas layanan informasi secara digital [1][2]. Salah satu bentuk penerapan teknologi tersebut adalah penggunaan *Chatbot* berbasis *Natural Language Processing* (NLP) sebagai media layanan informasi otomatis. *Chatbot* memungkinkan pengguna memperoleh informasi secara cepat tanpa harus berkomunikasi langsung dengan petugas administrasi [3].

MA Bahrul Ulum Tanjungjaya Subang yang berlokasi di Desa Patimban Kecamatan Pusakanagara Kabupaten Subang Provinsi Jawa Barat masih melakukan layanan informasi pendaftaran siswa baru secara konvensional melalui komunikasi langsung dan media brosur.

Kondisi tersebut menyebabkan calon siswa maupun orang tua sering mengalami keterlambatan dalam memperoleh informasi. Selain itu, pihak administrasi juga harus menjawab pertanyaan yang sama secara berulang sehingga pelayanan menjadi kurang efisien.

*Natural Language Processing* (NLP) merupakan cabang kecerdasan buatan yang digunakan untuk memproses dan memahami bahasa alami manusia [4]. Dalam pengembangan *Chatbot* berbasis pencarian (*retrieval-based*), metode TF-IDF digunakan untuk memberikan bobot pada kata, sedangkan *Cosine similarity* digunakan untuk menghitung tingkat kemiripan antara pertanyaan pengguna dan data FAQ yang tersedia [5][6].

Beberapa penelitian sebelumnya menunjukkan bahwa chatbot berbasis NLP cukup efektif digunakan pada layanan informasi pendidikan. Penerapan chatbot pada layanan PPDB mampu menghasilkan jawaban otomatis dengan tingkat kecocokan yang baik [7]. Penelitian lain menunjukkan bahwa chatbot berbasis NLP dapat memberikan respons otomatis sesuai dataset pertanyaan pengguna [8]. Selain itu, metode TF-IDF dan *Cosine similarity* juga telah diterapkan pada chatbot layanan akademik dengan tingkat akurasi yang cukup tinggi [9]. Kombinasi metode TF-IDF, *Cosine similarity*, dan *Naïve Bayes* juga terbukti mampu meningkatkan kualitas pencocokan pertanyaan pada chatbot akademik [10].

Meskipun penelitian sebelumnya menunjukkan bahwa chatbot berbasis NLP cukup efektif digunakan pada layanan informasi pendidikan, sebagian besar penelitian masih berfokus pada layanan akademik umum dan belum secara khusus diterapkan pada layanan informasi pendaftaran siswa baru di lingkungan madrasah.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk merancang dan membangun *Chatbot* layanan informasi pendaftaran siswa baru berbasis NLP menggunakan metode TF-IDF dan *Cosine similarity* pada MA Bahrul Ulum Tanjungjaya Subang. Sistem yang dikembangkan diharapkan mampu membantu pengguna memperoleh informasi secara cepat serta meningkatkan efisiensi layanan informasi madrasah.

## II. METODE

Penelitian ini menggunakan desain studi kasus dengan objek penelitian berupa *Chatbot* layanan informasi pendaftaran siswa baru di MA Bahrul Ulum Tanjungjaya Subang. Sistem *Chatbot* dikembangkan berbasis web menggunakan bahasa pemrograman Python dan framework Flask. Metode yang digunakan dalam penelitian ini adalah *Natural Language Processing* (NLP) dengan algoritma TF-IDF dan *Cosine similarity* untuk menentukan tingkat kemiripan pertanyaan pengguna terhadap data FAQ.

Dataset yang digunakan berupa *Frequently Asked Questions* (FAQ) yang diperoleh dari pihak administrasi madrasah. Dataset terdiri dari pasangan pertanyaan dan jawaban yang berkaitan dengan layanan informasi pendaftaran siswa baru, seperti informasi pendaftaran, persyaratan, daftar ulang, jurusan, lokasi madrasah, biaya pendidikan, ekstrakurikuler, dan kontak madrasah.

Tahapan penelitian dimulai dari pengumpulan data FAQ, *preprocessing* teks, pembobotan TF-IDF, perhitungan *Cosine similarity*, hingga pengujian sistem *Chatbot*. Tahapan *preprocessing* dilakukan untuk membersihkan dan menyeragamkan teks sebelum dilakukan proses perhitungan. Tahapan *preprocessing* meliputi *case folding*, *tokenizing*, *stopword removal*, dan *stemming* menggunakan *library Sastrawi*.

Setelah *preprocessing* dilakukan, sistem menghitung bobot setiap kata menggunakan metode TF-IDF. Metode TF-IDF digunakan untuk mengetahui tingkat kepentingan kata pada dokumen pertanyaan dan data FAQ. Persamaan TF-IDF ditunjukkan pada Persamaan (1).

$$TFIDF(t, d) = TF(t, d) \times IDF(t) \quad (1)$$

Keterangan:

$TF(t, d)$  = frekuensi kemunculan *term*  $t$  pada dokumen  $d$ .

$IDF(t)$  = nilai *inverse document frequency* dari *term*  $t$ .

$TFIDF(t, d)$  = bobot akhir *term*  $t$  pada dokumen  $d$

$t$  = *term* atau kata.

$d$  = dokumen FAQ.

Selanjutnya, perhitungan kemiripan dilakukan menggunakan metode *Cosine similarity* untuk menentukan tingkat kedekatan antara pertanyaan pengguna dan data FAQ. Persamaan *Cosine similarity* ditunjukkan pada Persamaan (2).

$$\text{Cosine Similarity} = \frac{\sum_{i=1}^n (A_i \times B_i)}{\sqrt{\sum_{i=1}^n A_i^2 \times \sum_{i=1}^n B_i^2}} \quad (2)$$

$A_i$  merupakan bobot TF-IDF pertanyaan pengguna

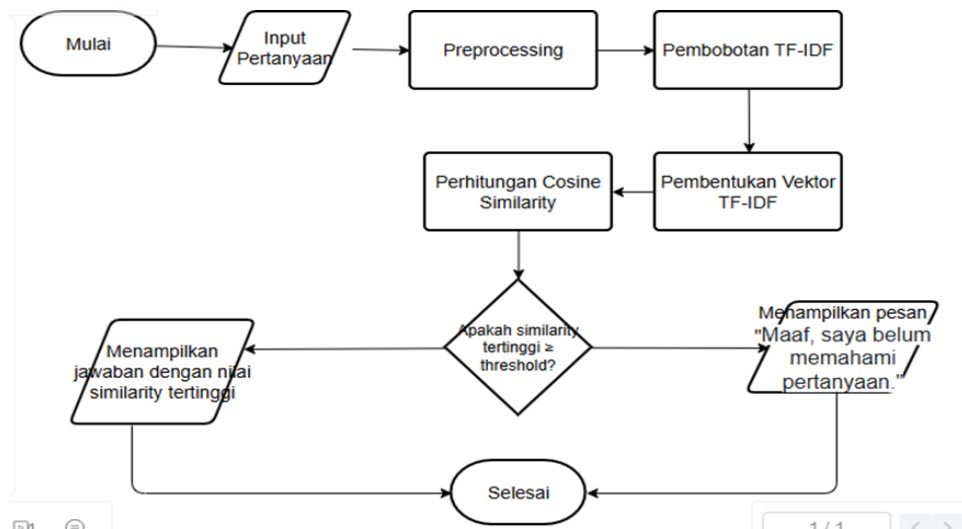
$B_i$  merupakan bobot TF-IDF dokumen FAQ

Hasil similarity berada pada rentang 0 hingga 1.

Sistem akan memilih jawaban dengan nilai similarity tertinggi. Apabila nilai similarity berada di bawah threshold 0,70 maka sistem akan menampilkan pesan bahwa informasi belum tersedia.

Pengujian sistem dilakukan menggunakan pengujian fungsional dan non fungsional. Pengujian fungsional dilakukan dengan memberikan beberapa variasi pertanyaan pada setiap kategori FAQ untuk mengetahui ketepatan jawaban *Chatbot*. Sedangkan pengujian non fungsional dilakukan untuk mengetahui kualitas sistem dari aspek kemudahan penggunaan, kecepatan respon, dan stabilitas sistem.

Flowchart system *Chatbot* dapat dilihat pada gambar 1.



Gambar 1. Flowchart Sistem *Chatbot*

Selain itu, struktur proses system *Chatbot* dalam penelitian ini ditunjukkan pada Tabel 1.

Tabel 1. Struktur proses *Chatbot*

| No | Struktur proses <i>chatbot</i> |                                                      |
|----|--------------------------------|------------------------------------------------------|
|    | Tahapan                        | Deskripsi                                            |
| 1  | Input pertanyaan               | Pengguna memasukkan pertanyaan                       |
| 2  | <i>Preprocessing</i>           | Case folding, tokenizing, stopword removal, stemming |
| 3  | TF-IDF                         | Pembobotan kata                                      |
| 4  | <i>Cosine similarity</i>       | Perhitungan kemiripan                                |
| 5  | Output                         | Menampilkan jawaban <i>Chatbot</i>                   |

### III. HASIL DAN PEMBAHASAN

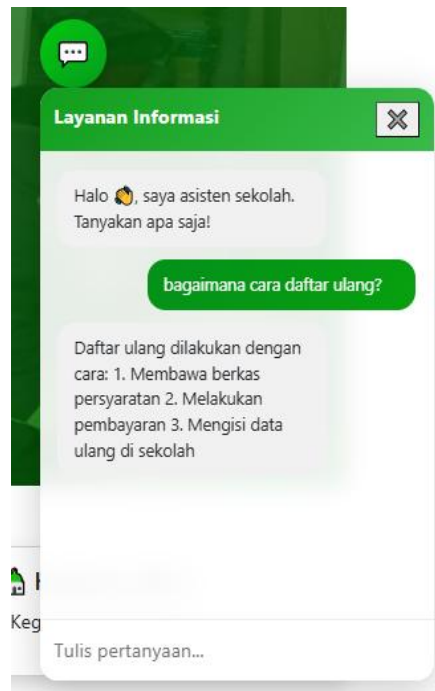
Sistem *Chatbot* layanan informasi pendaftaran siswa baru berhasil dibangun berbasis web menggunakan bahasa pemrograman Python dan framework Flask. Sistem ini dirancang untuk membantu pengguna memperoleh informasi pendaftaran secara otomatis melalui proses pencocokan pertanyaan menggunakan metode TF-IDF dan *Cosine similarity*.

#### A. Implementasi Sistem

Implementasi *Chatbot* terdiri dari beberapa modul utama, yaitu halaman *Chatbot*, modul *preprocessing* teks, modul pembobotan TF-IDF, modul perhitungan *Cosine similarity*, dan modul admin FAQ. Pengguna dapat memasukkan pertanyaan melalui halaman *Chatbot*, kemudian

sistem akan memproses pertanyaan tersebut untuk menentukan jawaban yang paling relevan berdasarkan dataset FAQ yang tersedia.

Antarmuka *Chatbot* digunakan sebagai media interaksi antara pengguna dan sistem. Pengguna dapat memasukkan pertanyaan terkait informasi pendaftaran siswa baru melalui halaman *Chatbot* seperti ditunjukkan pada Gambar 2.



Gambar 2. Tampilan Halaman *Chatbot*

Tahapan *preprocessing* yang diterapkan meliputi *case folding*, *tokenizing*, *stopword removal*, dan *stemming*. Tahapan ini bertujuan untuk membersihkan teks dan mengurangi kata yang tidak memiliki pengaruh besar terhadap proses pencocokan pertanyaan.

Contoh *preprocessing* ditunjukkan pada Tabel 2.

Tabel 2. Tahapan *Preprocessing*

| No | Tahapan preprocessing   |                                             |
|----|-------------------------|---------------------------------------------|
|    | Tahapan                 | Hasil                                       |
| 1  | <i>Input</i>            | Bagaimana cara daftar ulang siswa baru      |
| 2  | <i>Case folding</i>     | bagaimana cara daftar ulang siswa baru      |
| 3  | <i>Tokenizing</i>       | bagaimana, cara, daftar, ulang, siswa, baru |
| 4  | <i>Stopword removal</i> | cara, daftar, ulang, siswa, baru            |
| 5  | <i>Stemming</i>         | cara, daftar, ulang, siswa, baru            |

Setelah *preprocessing* selesai dilakukan, sistem menghitung bobot kata menggunakan tf-idf dan menentukan tingkat kemiripan menggunakan *Cosine similarity*.

## B. Hasil Pengujian

Pengujian sistem dilakukan menggunakan 80 pertanyaan uji yang dibagi ke dalam delapan kategori layanan informasi, yaitu pendaftaran, persyaratan, daftar ulang, jurusan, lokasi madrasah, ekstrakurikuler, biaya pendidikan, dan kontak madrasah.

Hasil pengujian ditunjukkan pada tabel 3.

Tabel 3. Hasil Pengujian

| NO | Hasil pengujian  |            |          |              |
|----|------------------|------------|----------|--------------|
|    | Kategori         | Jumlah uji | Berhasil | Tidak sesuai |
| 1  | Pendaftaran      | 10         | 7        | 3            |
| 2  | Persyaratan      | 10         | 9        | 1            |
| 3  | Daftar ulang     | 10         | 8        | 2            |
| 4  | Jurusan          | 10         | 8        | 2            |
| 5  | Lokasi madrasah  | 10         | 8        | 2            |
| 6  | Ekstrakurikuler  | 10         | 9        | 1            |
| 7  | Biaya pendidikan | 10         | 7        | 3            |
| 8  | Kontak madrasah  | 10         | 7        | 3            |
|    | Total            | 80         | 63       | 17           |

Berdasarkan hasil pengujian, *Chatbot* berhasil memberikan jawaban yang sesuai pada 63 pertanyaan dari total 80 pertanyaan uji. Tingkat keberhasilan sistem dihitung menggunakan Persamaan (3).

$$\text{Akurasi} = \frac{63}{80} \times 100\% = 78.75\%$$

Hasil tersebut menunjukkan bahwa metode TF-IDF dan *Cosine similarity* cukup efektif digunakan pada *Chatbot* layanan informasi berbasis *retrieval-based*.

Beberapa pertanyaan berhasil dijawab dengan baik meskipun memiliki variasi susunan kata yang berbeda. Sebagai contoh, pertanyaan “Mulai kapan siswa baru bisa daftar” dan “Kapan siswa baru bisa daftar” menghasilkan jawaban yang sesuai karena memiliki tingkat kemiripan yang tinggi terhadap data FAQ.

Namun demikian, masih terdapat beberapa pertanyaan yang menghasilkan jawaban tidak sesuai. Hal tersebut terjadi karena variasi kalimat pengguna belum seluruhnya tersedia dalam dataset FAQ. Selain itu, beberapa pertanyaan pendek memiliki nilai *similarity* yang rendah sehingga sistem tidak dapat menentukan jawaban yang tepat.

Penggunaan *threshold* sebesar 0,70 membantu sistem menghindari pemberian jawaban yang tidak relevan. Jika nilai *similarity* berada di bawah *threshold*, *Chatbot* akan menampilkan pesan bahwa informasi belum tersedia.

Hasil penelitian ini menunjukkan bahwa *Chatbot* mampu membantu proses layanan informasi pendaftaran siswa baru secara lebih cepat dibandingkan sistem manual sebelumnya. Sebelum implementasi *Chatbot*, pengguna harus menghubungi pihak administrasi secara langsung untuk

memperoleh informasi. Setelah *Chatbot* diterapkan, pengguna dapat memperoleh informasi secara otomatis melalui website tanpa bergantung pada petugas administrasi.

Penelitian ini juga sejalan dengan penelitian sebelumnya yang menunjukkan bahwa metode TF-IDF dan *Cosine similarity* efektif digunakan dalam *Chatbot* layanan informasi [3]–[5]. Perbedaannya, penelitian ini lebih difokuskan pada layanan informasi pendaftaran siswa baru di lingkungan madrasah.

Meskipun sistem telah berjalan dengan baik, *Chatbot* masih memiliki beberapa keterbatasan, yaitu belum mampu memahami seluruh variasi bahasa pengguna dan masih bergantung pada dataset FAQ yang tersedia. Oleh karena itu, penelitian selanjutnya dapat mengembangkan *Chatbot* menggunakan metode machine learning atau deep learning agar sistem mampu memahami konteks pertanyaan secara lebih baik.

#### IV. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, *Chatbot* layanan informasi pendaftaran siswa baru berbasis *Natural Language Processing* menggunakan metode TF-IDF dan *Cosine similarity* berhasil dirancang dan diimplementasikan pada MA Bahrul Ulum Tanjungjaya Subang. Sistem *Chatbot* mampu membantu pengguna memperoleh informasi pendaftaran secara otomatis melalui website.

Hasil pengujian menunjukkan bahwa *Chatbot* mampu memberikan jawaban yang sesuai pada 63 dari 80 pertanyaan uji dengan tingkat keberhasilan sebesar 78,75%. Metode TF-IDF berhasil digunakan untuk pembobotan kata, sedangkan *Cosine similarity* mampu menentukan tingkat kemiripan antara pertanyaan pengguna dan data FAQ sehingga sistem dapat memilih jawaban yang paling relevan.

Penerapan *Chatbot* juga membantu meningkatkan efisiensi layanan informasi dibandingkan proses manual sebelumnya karena pengguna dapat memperoleh informasi secara lebih cepat tanpa harus menghubungi pihak administrasi secara langsung.

Meskipun demikian, *Chatbot* masih memiliki keterbatasan dalam memahami variasi bahasa pengguna dan masih bergantung pada dataset FAQ yang tersedia. Oleh karena itu, pengembangan lebih lanjut dapat dilakukan dengan menambahkan dataset yang lebih bervariasi serta menerapkan metode machine learning atau deep learning agar sistem mampu memahami konteks pertanyaan secara lebih baik.

#### DAFTAR PUSTAKA

- [1] R. Rismawati, T. Ibrahim, I. Kartika, and O. Arifudin, “PERAN SISTEM INFORMASI DALAM MENINGKATKAN MUTU LAYANAN PENDIDIKAN,” *J. Tahsinia*, vol. Vol. 5, No, pp. 1099–1122, 2024, doi: <https://doi.org/10.57171/jt.v5i7.618>.
- [2] Nawafilillah, F. Zahroh, A. Z. Agustina, L. Cahyani, N. Aini, and E. Y. Wijaya, “PENGELOLAAN WEBSITE INFORMASI MADRASAH SEBAGAI MEDIA PROMOSI DIGITAL DI SMPN 1 LABANG,” *PAKDEMAS J. Pengabd. Kpd. Masy.*, vol. 4, no 2, 2025, doi: <https://doi.org/10.58222/pakdemas.v4i2.322>.
- [3] R. A. Sanjaya and E. Winarno, “Pengembangan Chatbot Informasi Pariwisata di Kabupaten Pati Menggunakan Metode Natural Language Processing Berbasis Dialogflow,” *Jutisi J. Ilm. Tek. Inform. dan Sist. Inf.*, vol. 13, No. 1, 2024, doi: <https://doi.org/10.35889/jutisi.v13i1.1828>.

- [4] S. Hidayat, A. Silvanie, and R. Subekti, "CHATBOT UNTUK KONSULTASI AKADEMIK MENGGUNAKAN NATURAL LANGUAGE PROCESSING(NLP) DI IBI-K 1957," DOI 10.52362/jisamar.v6i2.780 Ciptaan disebarluaskan di bawah Lisensi Creat. Commons Atribusi 4.0 Internasional. 396 *Journal Inf. Syst. Applied, Manag. Account. Res.*, vol. 6 No.2, 2022, doi: <https://doi.org/10.52362/jisamar.v6i2.780>.
- [5] Y. A. P. G. S, G. L. Ginting, and M. Panjaitan, "Optimasi Penerimaan Siswa Baru dengan Penerapan Algoritma Text Mining dan TF-IDF," *J. Comput. Informatics Res.*, vol. 2, No 3, 2023, doi: DOI 10.47065/comforch.v2i3.941.
- [6] D. A. A. Deepal, A. M. R. R. Bandara, and P. R. S. De Silva, "Siamese Hybrid Network Approach for Sentence Similarity," *Curr. Sci.*, vol. 7, No. 2, 2024, doi: <https://doi.org/10.31357/vjs.v27i02.7833>.
- [7] M. Al Farizi, I. N. Farida, and D. Swanjaya, "Penerapan Chatbot Berbasis Natural Language Processing untuk Layanan Informasi PPDB di SMK PGRI 1 Kediri," *SEMNAS INOTEK*, vol. 9, No. 2, 2025, doi: <https://doi.org/10.29407/c4x80271>.
- [8] R. R. Rafi, A. Setiawan, Asyadam, and A. Sanjaya, "CHATBOT PENERIMAAN PESERTA DIDIK BARU DI SMK INTENSIF BAITUSSALAM MENGGUNAKAN NLP," *Semin. Nas. Teknol. Sains*, vol. 4 No. 1, 2025, doi: <https://doi.org/10.29407/4vzbz3e19>.
- [9] N. Hikmah, D. Ariyanti, and F. A. Pratama, "Implementasi Chatbot Sebagai Virtual Assistant di Universitas Panca Marga Probolinggo menggunakan Metode TF-IDF," *J. Teknol. Inf. dan Multimed.*, vol. 4, No. 2, 2022, doi: <https://doi.org/10.35746/jtim.v4i2.225>.
- [10] R. N. Aziza, E. Yosrita, R. F. Ningrum, T. SukmaArdanti, and S. RafhilArafaizin, "Pembangunan Aplikasi dan Klasifikasi Pertanyaan Chatbot Informasi Akademik Menggunakan Metode Cosine Similarity dan Naïve Bayes," *KILAT*, vol. 2, No. 2, 2023, doi: <https://doi.org/10.33322/kilat.v12i2.1921>.