Journal of Applied Informatics and Computing (JAIC)
Vol.7, No.2, December 2023, pp. 252~257
e-ISSN: 2548-6861

Improving Helpdesk Chatbot Performance with Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity Models

Gede Herdian Setiawan [1]*, I Made Budi Adnyana [2]**
* Sistem Komputer, Institut Teknologi dan Bisnis STIKOM Bali
** Sistem Informasi, Institut Teknologi dan Bisnis STIKOM Bali
herdian@stikom-bali.ac.id [1], budi.adnyana@stikom-bali.ac.id [2]

Article Info

Article history:

Received 2023-09-12
Revised 2023-09-30
Accepted 2023-10-03

Keyword:

Chatbot,
TF-IDF,
Cosine Similarity,
NLP.

ABSTRACT

Helpdesk chatbots are growing in popularity due to their ability to provide help and answers to user questions quickly and effectively. Chatbot development poses several challenges, including enhancing accuracy in understanding user queries and providing relevant responses while improving problem-solving efficiency. In this research, we aim to enhance the accuracy and efficiency of the Helpdesk Chatbot by implementing the Term Frequency-Inverse Document Frequency (TF-IDF) model and the Cosine Similarity algorithm. The TF-IDF model is a method used to measure the frequency of words in a document and their occurrence in the entire document collection, while the Cosine Similarity algorithm is used to measure the similarity between two documents. After implementing and testing TF-IDF and Cosine Similarity models in the Helpdesk Chatbot, we achieved a 75% question recognition rate. To increase accuracy and precision, it is necessary to increase the knowledge dataset and improve pre-processing, especially in recognition and correct inaccurate spelling.

I. PENDAHULUAN

Chatbot semakin populer karena kemampuannya untuk memberikan bantuan dan jawaban atas pertanyaan pengguna dengan cepat dan efektif [1]. *Chatbot Helpdesk* khususnya adalah jenis *Chatbot* yang dirancang untuk memberikan dukungan pelanggan dengan menjawab pertanyaan seputar produk atau layanan [2]. *Chatbot* dikembangkan untuk dapat merespons masukan pengguna secara aktif [3].

Beberapa permasalahan yang dapat diselesaikan dengan dukungan *Chatbot Helpdesk* antara lain: dapat membantu mengurangi waktu respon yang dibutuhkan oleh layanan pelanggan dalam menanggapi pertanyaan atau masalah yang diajukan oleh pengguna[4] hal ini dapat membantu meningkatkan efisiensi layanan pelanggan dan mengurangi antrian yang terjadi, *Chatbot Helpdesk* dapat memberikan respons yang lebih cepat dan konsisten terhadap setiap pertanyaan atau masalah yang diajukan oleh pengguna sehingga dapat meningkatkan kualitas layanan pelanggan [5], *Chatbot Helpdesk* dapat mengurangi biaya operasional layanan pelanggan karena tidak memerlukan tenaga kerja manusia dalam menanggapi pertanyaan atau masalah pengguna [6], *Chatbot Helpdesk* yang responsif dan memberikan jawaban yang tepat dan relevan dapat meningkatkan kepuasan pengguna terhadap layanan pelanggan [7].

Dalam pengembangan *Chatbot*, terdapat beberapa tantangan yang harus diatasi, seperti peningkatan akurasi dalam memahami pertanyaan pengguna dan memberikan jawaban yang relevan. Untuk menganalisis teks dan memahami makna dari pertanyaan atau kalimat yang diajukan oleh pengguna salah satu model text mining yang bisa digunakan adalah *Term Frequency-Inverse Document Frequency* (TF-IDF) dan algoritma *Cosine Similarity*. Model TF-IDF adalah metode yang digunakan untuk mengukur frekuensi kata dalam dokumen dan kemunculannya di seluruh koleksi dokumen, sedangkan algoritma *Cosine Similarity* digunakan untuk mengukur kesamaan antara dua dokumen.

Beberapa penelitian yang telah menerapkan model TF-IDF dan *Cosine Similarity* seperti yang dilakukan oleh Defit S dkk meningkatkan akurasi dan pencarian kodefikasi barang hasilnya algoritma *Cosine Similarity* dan TF-IDF mampu

meningkatkan akurasi pencarian kodefikasi barang [8]. Chiny M mencari rekomendasi berdasarkan tren terikini pada Netflix hasilnya mampu meningkatkan probabilitas rekomendasi kepada pengguna [9]. Ristanti P melakukan klasifikasi jurnal ekonomi berdasarkan judul dan ringkasan hasilnya kinerja sistem menghasilkan akurasi tidak terlalu tinggi karena tidak melakukan pra-pemrosesan [10].

Berdasarkan penelitian terdahulu model TF-IDF dan *cosine similarity* terbukti mampu memberikan perbandingan antar dokumen teks dengan sangat baik. Pada penelitian ini model *Term Frequency-Inverse Document Frequency* (TF-IDF) dan algoritma *Cosine Similarity* digunakan untuk meningkatkan kemampuan *chatbot helpdesk* dalam mengukur kesamaan antara dua dokumen pada *dataset* jawaban dan pertanyaan dari pengguna. Penelitian sebelumnya penulis menerapakan *cosine similarity* dan BoW *(Bag of Word)* [11]. BoW memiliki keterbatasan dalam mempertimbangkan pentingnya bobot pada setiap kata, sehingga kata-kata yang kurang informatif dapat mempengaruhi dalam mengukur kesamaan.

Dengan menerapkan model TF-IDF untuk memperoleh representasi vektor dokumen dengan memperhatikan bobot lebih pada setiap kata dan algoritma *Cosine Similarity* untuk menghitung kemiripan antara vektor pertanyaan pengguna dan vektor dokumen jawaban, *Chatbot Helpdesk* pada penelitian ini diharapkan dapat memberikan respons yang lebih tepat dan relevan untuk setiap pertanyaan yang diajukan oleh pengguna.

II. METODE PENELITIAN

Metode atau tahapan penelitian dimulai dengan pengumpulan data. Dokumen atau data yang digunakan menjadi sumber data atau refresi dari *Chatbot*. Pada penelitian ini data yang digunakan adalah data dukungan teknis penggunaan sistem informasi pada perguruan tinggi. Data berupa pertanyaan yang sering diajukan oleh pengguna terkait penggunaan sistem informasi.

Tahapan pra-pemrosesan data dilakukan untuk membersihkan data teks dengan menghilangkan karakter-karakter yang tidak relevan dan melakukan stemming pada kata-kata pada dokumen.

Penerapan model TF-IDF pada dokumen, untuk mengukur frekuensi kata dalam dokumen dan kemunculannya di seluruh koleksi dokumen [12]. Rumus untuk TF-IDF sebagai berikut.

$$t\,f = 0{,}5 + 0{,}5\,x\,\frac{tf}{\max(tf)}$$

$$idf_t = \log(\frac{D}{df_t})$$

$$W_{d.t} = tf_{d.t}\,x\,idf_{d.t}$$

Dimana :

d	= dokumen ke-d
W	= term ke-t dari dokumen
tf	= banyaknya term I pada sebuah dokumen
idf	= inversed Document Frequency
df	= banyak dokumen yang mengandung term i

Selanjutnya penerapan algoritma *Cosine Similarity* untuk menghitung kemiripan antara vektor pertanyaan pengguna dan vektor dokumen jawaban. Rumus untuk algoritma *Cosine Similarity* sebagai berikut.

$$Cos\ a = \frac{\mathrm{A} \circ \mathrm{B}}{|\mathrm{A}|\,|\mathrm{B}|} = \frac{\sum_{i=1}^{n} \mathrm{A}_i\ x\ \mathrm{B}_i}{\sqrt{\sum_{i=1}^{n}(\mathrm{A}_i)^2\ x\ \sqrt{\sum_{i=1}^{n}(B_i)^2}}}$$

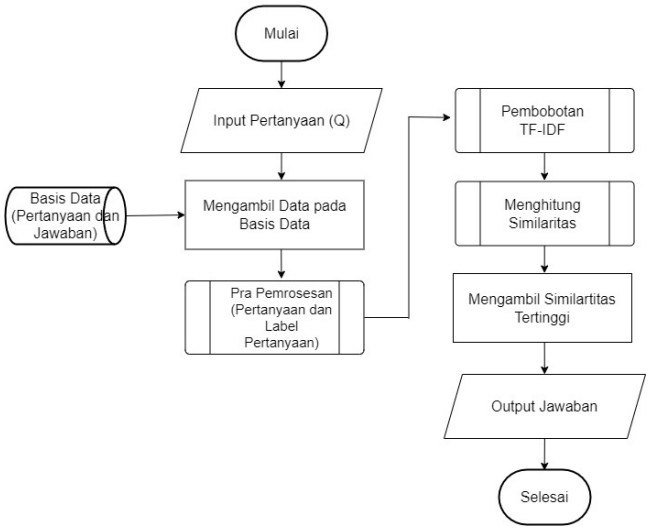
Dimana :

A	= vektor A yang akan dibandingkan
B	= vektor B yang akan dibandingkan
A o B	= dot product antara vector A dan vector B
|A|	= panjang vektor A
|B|	= panjang vektor B
|A||B|	= cross product antara |A| dan |B|

Evaluasi kinerja *Chatbot* dengan menggunakan metrik-metrik seperti precision, recall, dan F1-score. Precision mengukur seberapa akurat *Chatbot* dalam mengenali jawaban yang tepat. Precision didefinisikan sebagai rasio antara jumlah jawaban yang benar diidentifikasi oleh *Chatbot* dibandingkan dengan total jumlah jawaban yang diidentifikasi oleh *Chatbot*. Recall mengukur seberapa baik *Chatbot* dalam menemukan semua jawaban yang relevan. Recall didefinisikan sebagai rasio antara jumlah jawaban yang benar diidentifikasi oleh *Chatbot* dibandingkan dengan total jumlah jawaban yang seharusnya diidentifikasi oleh *Chatbot*. F1-score memberikan gambaran yang lebih baik tentang kinerja *Chatbot* secara keseluruhan karena mempertimbangkan baik precision maupun recall.

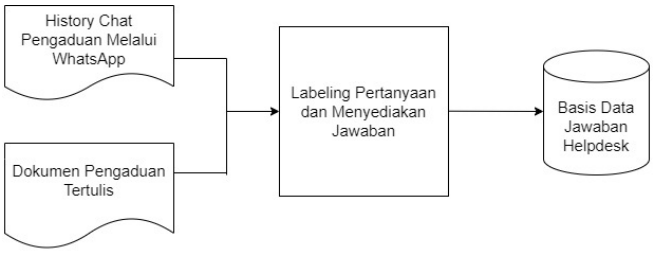
III. HASIL DAN PEMBAHASAN

Dalam meningkatkan akurasi untuk memahami pertanyaan pengguna dan memberikan jawaban yang relevan pada *Chatbot Helpdesk* penelitian ini menggunakan pendekatan Natural Language Processing (NLP) dengan menerapkan metode *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Cosine Similarity* [13]. Penerapan TF-IDF memiliki keuntungan dalam membantu mengenali kata-kata paling relevan dalam konteks dokumen tertentu dan dapat mengurangi kata-kata umum sehingga membantu meningkatkan kualitas perhitungan sedangkan metode *Cosine Similarity* dan membantu mengukur seberapa mirip dua vektor teks dalam ruang berdimensi tinggi pada teks pertanyaan pengguna dengan teks pertanyaan dalam basis data yang sudah disertakan dengan jawaban. Secara umum alur kerja *Chatbot Helpdesk* dengan pendekatan NLP seperti pada Gambar 1.



Gambar 1. Rancangan NLP pada Chatbot

Sebelum tahap pengembangan sistem berdasarkan uraian metode atau tahap penelitian pada pembahasan sebelumnya. Untuk dapat menyediakan basis data pertanyaan dan jawaban pada *Chatbot Helpdesk* terlebih dahulu dilakukan pengumpulan data pertanyaan berkaitan dengan dukungan teknis penggunaan sistem informasi yang diambil pada *history chat* mahasiswa pada layanan informasi akademik melalui platform whatsapp. Alur pengumpulan data dan pembentukan basis data ditujukan pada Gambar 2.



Gambar 2. Alur Pengumpulan Data

Seperti ditujukan pada Gambar 2 data pertanyaan dijadikan basis pengetahuan *Chatbot Helpdesk*. *Dataset* pertanyaan teknis yang terkumpul sampai saat ini berjumlah seratus pertanyaan, *dataset* pertanyaan ditujukan pada Tabel 1.

TABEL 1.
DATASET PERTANYAAN CHATBOT

No	Pertanyaan
1	Siapa saja yang mendapatkan akun email Microsoft?
2	Apa alamat akun email Microsoft untuk Mahasiswa?
3	Bagaimana mendapatkan password akun Microsoft untuk mahasiswa?
4	Bagaimana cara reset password Microsoft?
5	Saya sudah menunggu 5 menit, tapi tidak muncul di inbox email pribadi saya
.	……
96	Apa yang dilakukan jika tidak berhasil login pada e-krs?
97	Berapa jumlah SKS maksimal yang bisa diambil?
98	Apa saja yang harus dipersiapkan dalam perwalian?
99	Bagaimana cara menghubungi tim support perwalian?
100	Jika SKS saya belum maksimal, Bagaimana cara melakukan pemaksimalan?

Pada pra-pemrosesan data tahapan yang dilakukan sebagai berikut. Melakukan *Case Folding* untuk menyeragamkan karakter pada teks *dataset* dengan merubah seluruh huruf pada teks mejadi huruf kecil, menghapus tanda baca dan menghilangkan karakter-karakter yang tidak relevan. Melakukan tokenisasi bertujuan untuk memisahkan atau membagi teks kalimat menjadi kata-kata sebagai contoh pada kalimat pertanyaan *"bagaimana cara reset password microsoft"* akan menjadi *['bagaimana', 'cara', 'reset', 'password', 'microsoft']*. Selanjutnya dilakukan proses stemming untuk mengurai bentuk kata menjadi kata dasar, melakukan *stemming* atau lemmatisasi untuk mengubah kata-kata menjadi bentuk dasar dan menghilangkan kata-kata yang umum *(stop words removal)*. Setelah tahap pra-perosesan selesai selanjutnya melakukan penerapan model *Term Frequency-Inverse Document Frequency* (TF-IDF). TF-IDF digunakan untuk mengukur pentingnya sebuah kata dalam suatu dokumen dalam konteks keseluruhan korpus dokumen. Hal ini dilakukan dengan menggabungkan *term frequency* (TF), yaitu seberapa sering kata tersebut muncul dalam dokumen, dengan *inverse document frequency* (IDF), yaitu seberapa uniknya kata tersebut dalam korpus. Sebagai contoh pada penelitian ini pertanyaan : *"cara melakukan perwalian"* setelah melalui proses pra-pemrosesan *['wali']* dan kumpulan data pertanyaan yang telah diberikan label jawaban yang selanjutnya disebut dokumen seperti pada Tabel 2.

TABEL 2.
HASIL PRA-PEMROSESAN

Dokumen	Pra-Pemrosesan
D1	'akun', 'email', 'microsoft'
D2	'alamat', 'akun', 'email', 'microsoft'
D3	'password', 'akun', 'microsoft'
D4	'reset', 'password', 'microsoft'
D5	'tunggu', 'menit,', 'muncul', 'inbox', 'email', 'pribadi'
..	……
D96	'hubung', 'tim', 'support', 'perwalian'
D97	'sks', 'maksimal,', 'maksimal'
D98	'jadwal', 'wali', 'muncul'
D99	'terlabat', 'wali'
D100	'mana', 'daftar', 'matakuliah', 'wali'

Nilai TF-IDF untuk input pertanyaan terhadap seluruh dokumen setelah melalui proses perhitungan ditujukan pada Tabel 3.

TABEL 3.
NILAI TF-IDF

Nilai TF-IDF							
	D1	**D2**	**D3**	**--**	**D98**	**D99**	**D100**
akun	0.59	0.47	0.57	--	0	0	0
alamat	0	0.60	0	--	0	0	0
ambil	0	0	0	--	0	0	0
daftar	0	0	0	--	0	0	0.55
email	0.59	0.47	0	--	0	0	0
hasil	0	0	0	--	0	0	0
hubung	0	0	0	--	0	0	0
inbox	0	0	0	--	0	0	0
jadwal	0	0	0	--	0.69	0	0
krs	0	0	0	--	0	0	0
login	0	0	0	--	0	0	0
maksimal	0	0	0	--	0	0	0
mana	0	0	0	--	0	0	0.55
matakuliah	0	0	0	--	0	0	0.55
menit	0	0	0	--	0	0	0
microsoft	0.53	0.42	0.51	--	0	0	0
muncul	0	0	0	--	0.60	0	0
password	0	0	0.63	--	0	0	0
pribadi	0	0	0	--	0	0	0
reset	0	0	0	--	0	0	0
siap	0	0	0	--	0	0	0
sks	0	0	0	--	0	0	0
support	0	0	0	--	0	0	0
terlabat	0	0	0	--	0	0.87	0
tim	0	0	0	--	0	0	0
tunggu	0	0	0	--	0	0	0
wali	0	0	0	--	0.38	0.48	0.302

Setelah menghasilkan nilai/bobot TF-IDF pada pertanyaan pengguna dan seluruh dokumen *dataset* pada *database* berikutnya penerapan metode *cosine similirity* untuk mengukur seberapa mirip dua dokumen berdasarkan representasi vektor TF-IDF. Langkah penerapan *Cosine Similarity* sebagai berikut.

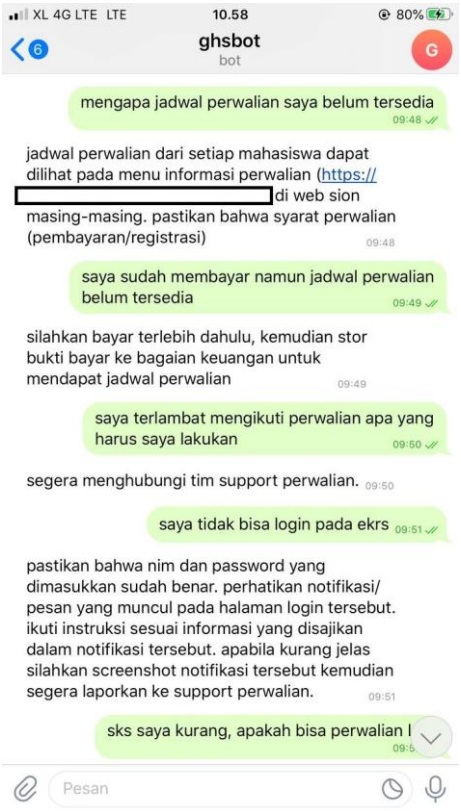
- Setiap dokumen (pertanyaan pengguna dan jawaban dalam *dataset*) direpresentasikan sebagai vektor dalam dimensi. Dimensi ruang setara dengan jumlah kata yang ada dalam seluruh dokumen dalam *dataset*.
- Perkalian antar vektor pertanyaan pengguna (A) dan vektor dokumen (B) dihitung dengan menjumlahkan hasil perkalian elemen-elemen vektor sehingga menghasilkan dot product.
- Menghitung jumlah vektor dengan menghitung akar kuadrat dari jumlah kuadrat nilai-nilai dalam vektor.
- Cosine Similarity antara dua vektor A dan B dihitung sebagai hasil bagi dari dot product antara vektor-vektor tersebut dan perkalian panjang vektor-vektor tersebut. Hasil Cosine Similarity berada dalam rentang -1 hingga 1. Nilai 1 menunjukkan bahwa vektor-vektor tersebut identik, nilai 0 menunjukkan bahwa vektor-vektor tersebut saling tegak lurus (tidak memiliki arah yang sama), dan nilai -1 menunjukkan bahwa vektor-vektor tersebut memiliki arah yang berlawanan.

Untuk mendapat jawaban yang paling relevan berdasarkan pertanyaan pengguna langkah perhitungan *Cosine Similarity* diulang untuk setiap dokumen dalam *dataset*. Hasil *Cosine Similarity* dihitung untuk setiap pasangan permintaan pengguna dan dokumen jawaban. Dokumen dengan nilai *Cosine Similarity* tertinggi dianggap sebagai *respons* yang paling relevan dan diberikan kepada pengguna [14].

TABEL 4.
HASIL PEMBOBOTAN COSINE SIMILARITY

No	**Dokumen *Dataset***	**Bobot Similarity**
1	D16	1
2	D22	0.8017837
3	D15	0.7071068
4	D23	0.7071068
5	D17	0.5
6	D18	0.5
7	D21	0.5
8	D20	0.3922323
9	D4	0.3535534
10	D27	0.3535534
11	D28	0.2886751
12	D24	0.2672612
13	D26	0.2672612

Tabel 4 menunjukan hasil pembobotan pertanyaan meggunakan *Cosine Similarity* untuk pertanyaan : *"cara melakukan perwalian"* ditemukan satu dokumen pertanyaan pada *dataset* yang memiliki nilai bobot similarity 1 atau sangat mirip dengan pertanyaan pada *dataset* sedangkan pada dua belas dokumen lainnya memiliki bobot lebih kecil dari 1. Jawaban atas dokumen pertanyaan yang memiliki bobot paling tinggi dijadikan respon / jawaban ke pengguna melalui *Chatbot*.



Gambar 3 Implementasi *chatbot*

Gambar 3. merupakan hasil implementasi model NLP dengan TF-IDF dan *Cosine similarity* pada chatbot helpdesk menunjukan chatbot telah berhasil memberikan respon jawaban susuai dengan *dataset* pengetahuan.

Setelah mendapatkan hasil selanjutnya dilakukan evaluasi dengan membandingkan hasil jawaban pada *Chatbot* dengan jawaban sebenarnya dalam *dataset*. Pada penelitian ini evaluasi dengan menghitung *accuracy, precision, recall dan F1-score*, dasar perhitungan mengacu pada norma *positive* dan *negative*. *Positve* ketika jawaban pada *Chatbot* dianggap relevan sedangkan negatif ketika jawaban dianggap tidak relevan dengan jawaban yang sebenarnya dalam *dataset*. Pada penelitian ini disiapkan dua puluh pertanyaan seputar topik yang sudah ada pada *dataset*. Untuk memudahkan perhitungan jawaban *Chatbot* diamati dan dikategorikan pada empat jenis jawaban yaitu : TP *(true positive)* ketika jawaban *Chatbot* benar dan sesuai dengan pertanyaan, TN *(true negative)* Ketika jawaban *Chatbot* tidak memberikan jawaban dan memang tidak ada dalam *dataset*, FP *(false positive)* ketika *Chatbot* menjawab namun jawaban dianggap keliru dan FN *(false negative) Chatbot* tidak memberikan jawaban namun sebenarnya ada pada *dataset*. Dua puluh pertanyaan dan hasil pengamatan jawaban ditujukkan pada tabel 5.

TABEL 5.
KETEGORI JAWABAN CHATBOT

No	Pertanyaan Uji	Kategori Jawaban
1	Bagaimana format password elearning?	TP
2	Saya lupa password elearning?	TP
3	Bagaimana cara reset password elearning?	FP
4	Saya lupa password ms teams?	FP
5	Bagaimana format password ms teams?	FP
6	Cara reset password ms teams?	TP
7	Saya sudah reset password namun belum masuk email?	TP
8	Jelaskan mengenai perwalian	FP
9	Bagaimana cara mengikuti perwalian?	FP
10	Apa saja syarat mengikuti perwalian?	TP
11	Dimana bisa melihat jadwal perwalian?	TP
12	Apa saja persiapan mengikuti perwalian?	TP
13	Mengapa jadwal perwalian saya belum tersedia?	TP
14	Saya sudah membayar namun jadwal perwalian belum tersedia	TP
15	Saya terlambat mengikuti perwalian apa yang harus saya lakukan?	TP
16	Saya tidak bisa login pada e-krs	TP
17	sks saya kurang, apakah bisa perwalian lagi?	FP
18	Bagaimana cara maksilkan sks	TP
19	Saya belum membayar, apakah bisa melakukan perwalian?	FP
20	Saya sudah selesai perwalian, selanjutnya bagaimana?	FP

Untuk menghitung performa *Chatbot* berdasarkan *accuracy, precision, recall dan F1-score* digunakan persamaan berikut.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} = 60\%$$

$$precision = \frac{TP}{TP + FP} = 60\%$$

$$recall = \frac{TP}{TP + FN} = 100\%$$

$$F1 - score = 2\,x\,\frac{recall\ x\ precision}{recall + precision} = 75\%$$

Nilai *accuracy* 60%: mengindikasikan bahwa *Chatbot* memiliki tingkat keberhasilan sekitar 60% dalam mengklasifikasikan pertanyaan atau permintaan dari pengguna dengan benar. Akurasi mengukur sejauh mana model dapat memprediksi dengan benar pada seluruh *dataset* pengujian. Hasil pengujian *precision* 60% menujukan berapa banyak jawaban positif yang diberikan oleh *Chatbot* yang relevan. *F1-score* memperoleh hasil 75% menunjukkan *Chatbot* memiliki performa yang cukup baik dalam menghasilkan respon jawaban yang relevan. *Chatbot* memiliki keseimbangan yang baik antara *precision* dan *recall* dan cendrung mampu memberikan respon yang tepat dan relevan.

IV. KESIMPULAN

Berdasarkan implementasi dan pengujian penerapan model TF-IDF dan *Cosine Similarity* pada *Chatbot Helpdesk* dapat ditarik kesimpulan bahwa *Chatbot* dengan cukup baik mampu mengenali pertanyaan dari pengguna dan mampu memberikan respon jawaban berdasarkan *dataset* dengan tingkat akurasi 60%, presisi 60% dan f1-score 75%. Namun demikian masih terdapat ruang untuk perbaikan terutama dalam meningkatkan akurasi dan presisi dengan mengumpulkan lebih banyak *dataset* pengetahuan, meningkatkan pra-pemrosesaan terutama dalam pengenalan dan perbaikan ejaan yang kurang tepat.

UCAPAN TERIMA KASIH

Ucapan terima kasih penulis sampaikan kepada Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi dan Institut Teknologi dan Bisnis STIKOM Bali yang telah mendukung peneliti baik dari segi moril maupun finansial dalam menyelesaikan penelitian ini.

DAFTAR PUSTAKA

[1] A. L. Chiru, I. A. Awada, and A. M. Florea, "A Support Process of Telemedicine Applications that Integrates a Chatbot," in *2021 International Conference on e-Health and Bioengineering (EHB)*, 2021, pp. 1–4. doi: 10.1109/EHB52898.2021.9657553.

[2] R. Shah, S. Lahoti, and K. Lavanya, "An intelligent chat-bot using natural language processing," *International Journal of Engineering Research*, vol. 6, no. 5, p. 281, 2017, doi: 10.5958/2319-6890.2017.00019.8.

[3] S. K. Maher, S. G. Bhable, A. R. Lahase, and S. S. Nimbhore, "AI and Deep Learning-driven Chatbots: A Comprehensive Analysis and Application Trends," in *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2022, pp. 994–998. doi: 10.1109/ICICCS53718.2022.9788276.

[4] J. J. Sophia and T. P. Jacob, "EDUBOT-A Chatbot For Education in Covid-19 Pandemic and VQAbot Comparison," in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2021, pp. 1707–1714. doi: 10.1109/ICESC51422.2021.9532611.

[5] P. D. Larasati, A. Irawan, S. Anwar, M. F. Mulya, M. A. Dewi, and I. Nurfatima, "Chatbot helpdesk design for digital customer service," *Applied Engineering and Technology*, vol. 1, no. 3, pp. 138–145, 2022, doi: 10.3176/aet.v1i1.684.

[6] D. C. Ukpabi, B. Aslam, and H. Karjaluoto, "Chatbot adoption in tourism services: A conceptual exploration," in *Robots, Artificial Intelligence and Service Automation in Travel, Tourism and Hospitality*, Emerald Group Publishing Ltd., 2019, pp. 105–121. doi: 10.1108/978-1-78756-687-320191006.

[7] A. Ali and M. Zain Amin, "Conversational AI Chatbot Based on Encoder-Decoder Architectures with Attention Mechanism," *Artificial Intelligence Festival*, vol. 2, no. 0, 2019, doi: 10.13140/RG.2.2.12710.27204.

[8] S. Defit and G. Widi Nurcahyo, "Product Codefication Accuracy With Cosine Similarity And Weighted Term Frequency And Inverse Document FREQUENCY (TF-IDF)," 2021.

[9] M. Chiny, M. Chihab, O. Bencharef, and Y. Chihab, "Netflix Recommendation System based on TF-IDF and Cosine Similarity Algorithms," Scitepress, May 2022, pp. 15–20. doi: 10.5220/0010727500003101.

[10] P. Y. Ristanti, A. P. Wibawa, and U. Pujianto, "Cosine Similarity for Title and Abstract of Economic Journal Classification," in *Proceeding - 2019 5th International Conference on Science in Information Technology: Embracing Industry 4.0: Towards Innovation in Cyber Physical System, ICSITech 2019*, Institute of Electrical and Electronics Engineers Inc., Oct. 2019, pp. 123–127. doi: 10.1109/ICSITech46713.2019.8987547.

[11] G. Herdian Setiawan and I. Made Budi Adnyana, "Information Retrieval Pada Frequently Asked Questions (FAQ) dengan metode String Similarity Information Retrieval on Frequently Asked Questions (FAQ) using String Similarity method," 2022.

[12] R. T. Wahyuni, D. Prastiyanto, and D. E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi."

[13] S. Ayanouz, B. A. Abdelhakim, and M. Benhmed, "A Smart Chatbot Architecture based NLP and Machine Learning for Health Care Assistance," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Mar. 2020. doi: 10.1145/3386723.3387897.

[14] Q. Xu, "Research on Text Classification Method based on PTF-IDF and Cosine Similarity," *Journal of Information and Communication Engineering*, vol. 6, no. 1, pp. 335–339, 2020, [Online]. Available: https://www.kaggle.com/shineucc/bbc-newsdataset