ANALISIS PENGELOMPOKKAN INTENT PERCAKAPAN CHATBOT MENGGUNAKAN METODE TF-IDF DAN ALGORITMA K-MEANS

Muhamad Ramadhan Heru Pradipta, Apriansyah Wahyudi Putra, Ramanuddin Jaya, Windah Kurnia Sari, Ken Ditha Tania, Fathoni
Fakultas Ilmu Komputer, Program Studi Sistem Informasi, Universitas Sriwijaya
Jl. Raya Palembang - Prabumulih No.KM. 32 Indralaya, Indonesia
09031282328030@student.unsri.ac.id

ABSTRAK

Penggunaan *chatbot* yang kian masif saat ini menghasilkan lonjakan volume data *log* percakapan pengguna secara signifikan. Namun, hal ini memunculkan permasalahan berupa kesulitan dalam memetakan maksud (*intent*) pengguna yang sangat beragam dan tidak terstruktur jika proses evaluasi dilakukan secara manual. Oleh karena itu, penelitian ini bertujuan untuk menganalisis pengelompokan *intent* secara otomatis pada dataset percakapan *chatbot multi-turn* untuk menemukan pola topik yang dominan. Metode penyelesaian yang digunakan berlandaskan kerangka *Knowledge Discovery in Database* (KDD), meliputi transformasi teks dengan *Term Frequency-Inverse Document Frequency* (TF-IDF), pengelompokan data menggunakan algoritma *K-Means Clustering*, serta validasi *cluster* dengan *Elbow Method* dan *Silhouette Score*. Hasil penelitian menunjukkan pembentukan *cluster* optimal pada K=54 dengan *Silhouette Score* tertinggi sebesar 0,5353. Kombinasi metode ini terbukti efektif dan granular dalam memisahkan spektrum kebutuhan pengguna mulai dari edukasi teknis hingga hiburan sehingga dapat menjadi landasan evaluasi bagi pengembang *chatbot* ke depannya.

Kata kunci : *Pengelompokkan, Intent, Percakapan, Chatbot AI, TF-IDF, K-Means*

1. PENDAHULUAN

Perkembangan teknologi berjalan sangat pesat di era revolusi industri sekarang ini. Kecerdasan buatan (*Artificial Intelligence*) salah satu contoh nyata perkembangan ini telah membawa perubahan signifikan dalam cara manusia berinteraksi dengan sistem komputer terutama penggunaan *chatbot* [1].

Chatbot atau sistem percakapan otomatis kini banyak diimplementasikan di berbagai sektor, mulai dari layanan pelanggan (*customer service*), pendidikan, hingga asisten virtual pribadi. Seiring dengan meningkatnya penggunaan *chatbot*, volume data percakapan antara pengguna dan sistem juga mengalami lonjakan yang masif [2].

Data percakapan ini menyimpan informasi berharga mengenai perilaku, kebutuhan, dan maksud (*intent*) pengguna yang sebenarnya [3].

Tantangan utama dalam pengelolaan *chatbot* adalah memahami variasi maksud (*intent*) dari pengguna yang sangat beragam dan tidak terstruktur. Seringkali, pengembang dari chatbot kesulitan untuk memetakan topik apa saja yang sering ditanyakan oleh pengguna karena data *log* percakapan yang masuk jumlahnya bisa mencapai jutaan percakapan dalam hitungan hari. Metode analisis manual tentu tidak lagi efektif dan memakan waktu yang sangat lama. Oleh karena itu , diperlukan pendekatan otomatis berbasis *Text Mining* untuk menggali informasi tersembunyi dari tumpukan data teks tersebut tanpa harus melabeli data satu per satu [4].

Penelitian ini bertujuan untuk menganalisis pengelompokan *intent* pada dataset percakapan *chatbot multi-turn* menggunakan kombinasi metode ekstraksi fitur TF-IDF dan algoritma *K-Means Clustering.* Dengan menerapkan metode ini, diharapkan pola-pola topik percakapan yang dominan dapat teridentifikasi secara otomatis. Hasil dari pengelompokan ini nantinya dapat memberikan wawasan bagi pengembang sistem untuk memahami intensi pengguna secara lebih mendalam dan meningkatkan kualitas respon *chatbot* di masa mendatang.

Salah satu solusi untuk mengatasi permasalahan ini adalah dengan menerapkan teknik *Clustering* atau pengelompokkan data. *CLustering* termasuk dalam kategori *Unsupervised Learning* yang memungkinkan sistem untuk mengelompokkan data berdasarkan kemiripan karakteristiknya tanpa memerlukan label kelas sebelumnya. Algoritma *K-Means* merupakan metode yang populer digunakan untuk kasus ini karena komputasinya yang efisien dan kemampuannya dalam menangani data dalam jumlah besar [7]. Supaya algoritma *K-Means* dapat memproses data teks, diperlukan proses ekstraksi fitur untuk mengubah data teks menjadi representasi dalam bentuk numerik. Metode *Term Frequency-Inverse Document Frequency* (TF-IDF) menjadi pilihan karena kemampuannya dalam mengkonversi kata-kata penting menjadi nilai numerik yang menjadi pembeda antar topik [8].

2. TINJAUAN PUSTAKA

2.1. *Text Mining* dan *Preprocessing*

Text mining merupakan salah satu cabang dari *data mining* yang secara khusus berfokus pada proses penemuan informasi, pola, atau pengetahuan baru yang berharga dari sekumpulan data berbentuk teks yang tidak terstruktur. Dalam pengolahan data teks percakapan *chatbot*, teks mentah tidak dapat langsung diproses oleh algoritma matematis. Oleh karena itu,

diperlukan tahapan *preprocessing* untuk membersihkan dan menstandarisasi data. Tahapan ini umumnya meliputi *case folding* (mengubah seluruh huruf menjadi huruf kecil), *tokenizing* (memecah kalimat menjadi potongan kata), *filtering* atau *stopword removal* (menghilangkan kata hubung yang tidak memiliki makna signifikan), dan *stemming* (mengembalikan kata ke bentuk dasarnya) [9].

2.2. *Chatbot* dan *Intent* Pengguna

Chatbot adalah program komputer berbasis kecerdasan buatan yang dirancang untuk menyimulasikan percakapan intelektual dengan satu atau lebih pengguna manusia baik melalui audio maupun teks. Kunci utama dari keberhasilan sebuah *chatbot* adalah kemampuannya dalam mengenali *intent* (maksud) dari input pengguna. *Intent* merepresentasikan tujuan atau apa yang sebenarnya ingin dicapai oleh pengguna ketika mengetikkan sebuah kalimat. Menganalisis *intent* dalam skala besar sangat penting untuk mengetahui tren topik keluhan atau kebutuhan pengguna secara komprehensif tanpa harus mengevaluasi percakapan secara manual [10].

2.3. *Term Frequency-Inverse Document Frequency* (TF-IDF)

TF-IDF adalah salah satu metode pembobotan kata yang paling umum digunakan dalam *Information Retrieval* dan *Text Mining*. Metode ini digunakan untuk mengevaluasi seberapa penting sebuah kata di dalam suatu dokumen yang merupakan bagian dari sebuah korpus (kumpulan dokumen teks). Nilai TF (*Term Frequency*) menghitung frekuensi kemunculan suatu kata dalam satu kalimat percakapan, sedangkan IDF (*Inverse Document Frequency*) mengurangi bobot kata-kata yang terlalu sering muncul di seluruh dataset percakapan. Melalui kombinasi ini, TF-IDF mampu mengonversi teks percakapan *chatbot* menjadi representasi matriks numerik yang siap dikelompokkan oleh algoritma *clustering* [11].

2.4. *K-Means Clustering*

K-Means merupakan salah satu algoritma *clustering* non-hierarki (*Unsupervised Learning*) yang mengelompokkan data ke dalam sejumlah *cluster* (K) berdasarkan jarak terdekat terhadap pusat *cluster* atau *centroid*. Algoritma ini bekerja secara iteratif untuk meminimalkan variasi antar data di dalam satu *cluster* yang sama, dan memaksimalkan variasi dengan data di *cluster* yang berbeda. Metode ini banyak digunakan dalam penelitian pengelompokan teks karena mampu menangani data dalam skala besar secara efisien dan menghasilkan pengelompokan yang mudah diinterpretasikan [12].

2.5. Evaluasi Kualitas *Cluster*

Evaluasi kualitas *cluster* diperlukan untuk memastikan bahwa hasil pengelompokan topik percakapan yang dihasilkan memiliki tingkat pemisahan dan homogenitas yang baik. *Elbow Method* digunakan untuk menentukan jumlah *cluster* optimal dengan menganalisis perubahan nilai *Within-Cluster Sum of Squares* (WCSS) terhadap penambahan jumlah *cluster*. Selain itu, pengujian *Silhouette Score* digunakan untuk mengukur tingkat kedekatan data teks dalam satu *cluster* dibandingkan dengan *cluster* lainnya, sehingga kualitas pengelompokan *intent* dapat divalidasi secara kuantitatif [6].

2.6. Penelitian Terdahulu

Sejumlah studi telah dilakukan terkait penerapan clustering pada data teks. Penelitian oleh Setiawan dkk. (2025) menggunakan kombinasi model IndoBERT dan K-Means untuk mengelompokkan topik keluhan mahasiswa berdasarkan makna semantik, yang menunjukkan bahwa K-Means mampu memisahkan topik-topik permasalahan secara efektif [5].

Penelitian lain oleh Agustine dkk. (2025) menerapkan algoritma K-Means untuk mengelompokkan wilayah berdasarkan aspek sosial ekonomi, di mana penggunaan prapemrosesan dan standarisasi data terbukti meningkatkan akurasi clustering secara signifikan [6].

Selain itu, penelitian terkait analisis data teks percakapan menggunakan K-Means dan TF-IDF juga pernah dilakukan untuk mengelompokkan ulasan pengguna pada platform digital, yang membuktikan bahwa TF-IDF adalah metode ekstraksi fitur yang efisien untuk algoritma berbasis jarak seperti K-Means. Berbeda dengan penelitian-penelitian terdahulu yang mayoritas berfokus pada analisis sentimen ulasan atau keluhan layanan publik, penelitian ini secara spesifik berfokus pada pengelompokan intent percakapan pada dataset chatbot multi-turn. Hal ini memberikan kebaruan (novelty) dalam melihat pola alur pertanyaan pengguna yang sangat dinamis untuk mengoptimalkan pemetaan topik pada sistem chatbot modern.

3. METODE PENELITIAN

3.1. Data

Sumber data dalam penelitian diambil melalui website Kaggle yaitu di *https://www.kaggle.com/datasets/abhayayare/multi-turn-chatbot-conversation-dataset.* Dataset ini merupakan kumpulanm percakapan multi-turn antara pengguna dan sistem chatbot berbasis *Large Language Model (LLM)* yang mencakup beragam topik pertanyaan dan permintaan informasi dari pengguna nyata.

Dataset ini merupakan kumpulan percakapan multi-turn antara pengguna dan sistem chatbot berbasis model bahasa besar (LLM) yang mencakup beragam topik pertanyaan dan permintaan informasi dari pengguna nyata [13].

Dataset ini dipilih karena memiliki beberapa keunggulan yang relevan dengan tujuan penelitian, antara lain :

(1) volume data yang besar sehingga representatif

untuk analisis pengelompokan berskala besar (2) variasi topik percakapan yang sangat beragam mencakup domain teknologi, kesehatan, pendidikan, hiburan, dan lainnya.

3.2. Tahapan Perancangan

Pada tahapan perancangan penelitian ini menggunakan metode perancangan KDD. Metode ini dipilih karena mampu mengakomodasi seluruh proses analisis data teks secara menyeluruh, mulai dari pemilihan data hingga diperolehnya pengetahuan yang dapat dimanfaatkan [14].

Tahapan KDD yang diterapkan dalam penelitian ini meliputi Selection Data, Preprocessing Data, Transformation Data, Data Mining, Evaluation Data, dan Knowledge.

Berikut ini merupakan tahapan dari proses *Knowledge Discovery in Database (KDD) :*

3.3. *Data Selection*

Tahap *selection* merupakan tahap pertama dalam kerangka KDD, yaitu proses pemilihan data yang relevan dari keseluruhan dataset yang tersedia. Pada tahap ini, dari keseluruhan kolom yang terdapat dalam dataset *ChatGPT-Style 1M Conversation Dataset (Multi-Turn)*, dipilih kolom *message* sebagai satu-satunya atribut yang akan dianalisis. Pemilihan kolom ini didasarkan pada pertimbangan bahwa kolom *message* secara langsung merepresentasikan input atau pertanyaan yang disampaikan pengguna kepada chatbot, sehingga paling relevan untuk mengidentifikasi intent pengguna.

3.4. Preprocessing Data

Tahap *preprocessing* bertujuan untuk membersihkan dan menstandarisasi data teks percakapan agar siap diolah oleh algoritma machine learning. Data teks mentah dari percakapan chatbot seringkali mengandung noise, inkonsistensi penulisan, dan karakter-karakter yang tidak relevan yang dapat mengganggu proses ekstraksi fitur.

3.5. *Transformation Data*

kata	jumlah_dokumen	skor_tfidf
tips	11794	0.051427
best	9943	0.0421282
resume	2616	0.0182016
explain	3213	0.0176953
suggest	2723	0.0175374

Gambar 1. TF-IDF Score

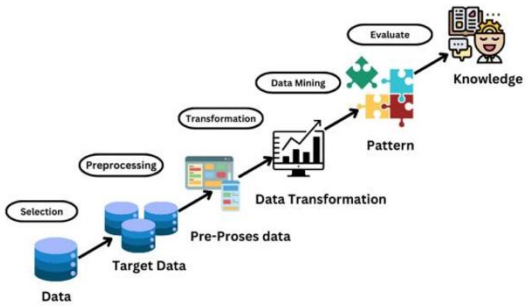
Pada tahap ini, data teks yang telah melalui *preprocessing* diubah menjadi representasi numerik menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF) [15]. Hasil transformasi menunjukkan bahwa kata tips memiliki skor TF-IDF tertinggi sebesar 0,051427 dengan kemunculan pada 11.794 dokumen, diikuti kata best (0,042128), resume (0,018202), explain (0,017695), dan suggest (0,017537). Nilai skor yang lebih rendah mengindikasikan bahwa kata tersebut tersebar lebih merata di seluruh korpus, sehingga memiliki daya diskriminatif yang lebih kecil sebagai fitur masukan algoritma *machine learning*.

3.6. *Data Mining*

Tahap *data mining* merupakan inti dari proses KDD, yaitu penerapan algoritma untuk menemukan pola tersembunyi dalam data. Pada penelitian ini, algoritma yang digunakan adalah *K-Means Clustering*, sebuah metode *unsupervised learning* yang mengelompokkan data ke dalam K cluster berdasarkan kemiripan jarak Euclidean terhadap centroid masing-masing cluster [16].

3.7. *Evaluation Data*

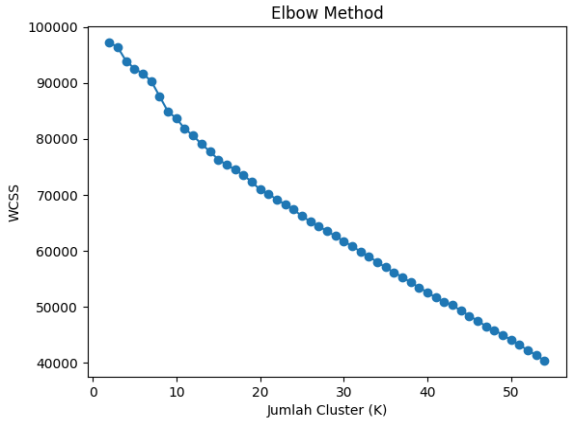
Tahap evaluation bertujuan untuk menilai kualitas hasil pengelompokan yang dihasilkan oleh algoritma K-Means. Pada penelitian ini, evaluasi dilakukan menggunakan dua metode kuantitatif, yaitu *Elbow Method* dan *Silhouette Score* [17].



Gambar 2. Tahapan perancangan

4. HASIL DAN PEMBAHASAN

4.1. Elbow Method



Gambar 3. Elbow Method

Pada Gambar 3 terlihat hubungan antara jumlah cluster (K) dan nilai WCSS menggunakan metode Elbow. Dari grafik tersebut, nilai WCSS memang terus menurun setiap kali jumlah cluster ditambah. Namun, penurunannya terlihat cukup stabil dan tidak menunjukkan adanya titik siku yang benar-benar jelas sebagai penanda jumlah cluster optimal. Artinya, sulit untuk menentukan secara pasti di K berapa pembagian cluster sebaiknya dilakukan hanya berdasarkan grafik ini, karena tidak ada perubahan penurunan yang sangat signifikan pada titik tertentu. Oleh sebab itu, untuk memastikan jumlah cluster yang dipilih lebih tepat dan tidak hanya berdasarkan visual grafik semata, hasil ini akan divalidasi kembali pada bagian berikutnya menggunakan Silhouette Score [18].

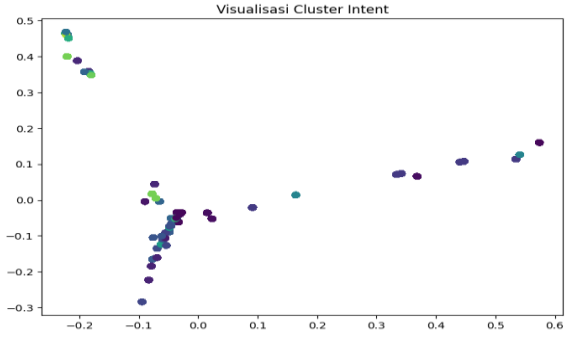
4.2. *Silhoutte Score*

```
K=43 | Silhouette=0.4347
K=44 | Silhouette=0.4441
K=45 | Silhouette=0.4535
K=46 | Silhouette=0.4626
K=47 | Silhouette=0.4722
K=48 | Silhouette=0.4804
K=49 | Silhouette=0.4892
K=50 | Silhouette=0.4983
K=51 | Silhouette=0.5073
K=52 | Silhouette=0.5165
K=53 | Silhouette=0.5255
K=54 | Silhouette=0.5353
```

Gambar 4. Silhouette Score

Gambar 4 menampilkan hasil evaluasi jumlah cluster menggunakan Silhouette Score pada beberapa nilai K. Berdasarkan hasil perhitungan, terlihat bahwa nilai Silhouette Score cenderung meningkat seiring dengan bertambahnya jumlah cluster. Nilai tertinggi diperoleh pada K = 54 dengan skor sebesar 0,5353, yang menunjukkan bahwa struktur cluster pada jumlah tersebut memiliki tingkat pemisahan antar cluster yang cukup baik serta kekompakan dalam cluster yang relatif optimal. Peningkatan skor yang konsisten ini mengindikasikan bahwa pembagian data menjadi 54 cluster memberikan kualitas pengelompokan yang lebih baik dibandingkan jumlah cluster yang lebih kecil. Oleh karena itu, berdasarkan evaluasi menggunakan Silhouette Score, jumlah cluster yang dipilih dalam penelitian ini adalah K = 54.

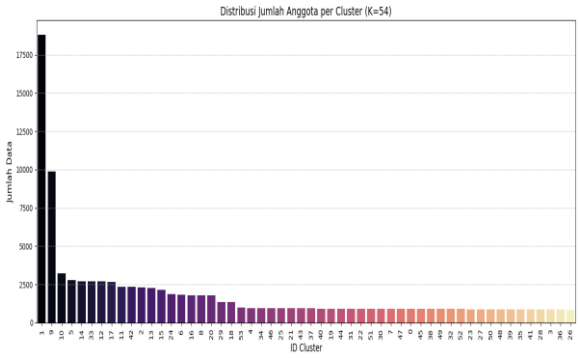
4.3. Visualisasi Hasil PCA



Gambar 5. Visualisasi Hasil PCA

Gambar 5 menampilkan visualisasi hasil clustering dengan K = 54 setelah dilakukan reduksi dimensi menggunakan Principal Component Analysis (PCA) ke dalam dua komponen utama pertama [19]. Setiap titik merepresentasikan satu data yang telah diproyeksikan dari ruang berdimensi tinggi ke ruang dua dimensi, dengan warna yang menunjukkan label cluster masing-masing. Berdasarkan visualisasi tersebut, terlihat bahwa beberapa cluster membentuk kelompok yang relatif terpisah, meskipun sebagian lainnya masih berada pada area yang berdekatan. Hal ini menunjukkan bahwa model K-Means mampu menangkap pola segmentasi dalam data yang dianalisis.

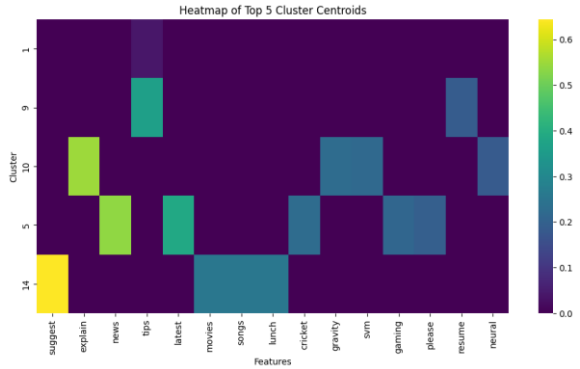
4.4. *Cluster*



Gambar 6. Distribusi Ukuran Cluster

Gambar 6 menunjukkan distribusi jumlah anggota pada setiap cluster dari total K = 54. Terlihat adanya ketimpangan populasi yang signifikan, di mana *Cluster* 1 dan *Cluster* 9 muncul sebagai kelompok mayoritas. Secara kumulatif, lima cluster terbesar (*Cluster* 1, 9, 10, 5, dan 14) mencakup sebagian besar populasi dataset. Distribusi ini memberikan landasan metodologis untuk memfokuskan analisis mendalam pada lima cluster dominan tersebut, karena dianggap paling representatif dalam menggambarkan tren utama perilaku data secara keseluruhan.

4.5. *Heatmap*



Gambar 7. Heatmap of Centroid Features

Gambar 7 menampilkan nilai rata-rata (*centroid*) fitur pada lima cluster terbesar untuk menginterpretasikan karakteristik dominan di tengah

jumlah cluster yang kompleks (K = 54). Cluster 10 menunjukkan intensitas tertinggi pada fitur *explain*, *gravity*, dan *neural* yang mengindikasikan fokus pada topik edukasi teknis, sementara Cluster 5 didominasi oleh fitur *news* dan *latest* yang mencerminkan ketertarikan pada informasi terkini. Cluster 14 menonjol dalam aspek rekomendasi melalui fitur *suggest* dan *movies*, sedangkan Cluster 9 lebih berorientasi pada pengembangan karier dengan fitur *tips* dan *resume*. Terakhir, Cluster 1 menunjukkan pola interaksi yang lebih umum dengan intensitas fitur yang relatif rendah, sehingga kelima cluster ini secara kolektif merepresentasikan spektrum kebutuhan utama pengguna mulai dari edukasi, berita, hingga hiburan.

4.6. Pembahasan

Penelitian ini berhasil mengidentifikasi struktur segmentasi intent percakapan pengguna yang kompleks melalui pendekatan dua tahap: penentuan jumlah cluster optimal (K = 54) dan analisis karakteristik mendalam pada lima cluster dominan. Tingginya nilai K yang divalidasi melalui Silhouette Score menunjukkan bahwa data memiliki granularitas yang tinggi, di mana intent pengguna tidak hanya terpusat pada satu topik besar, melainkan terfragmentasi ke dalam banyak segmen spesifik [10]. Hal ini mempertegas bahwa pendekatan segmentasi tunggal tidak lagi memadai untuk menangkap keberagaman perilaku pengguna dalam dataset ini.

Fenomena long-tail distribution yang terlihat pada distribusi ukuran cluster (Gambar 5) mencerminkan adanya polaritas minat yang sangat kontras di antara pengguna. Dominasi signifikan pada Cluster 1 dan Cluster 9 menunjukkan bahwa fungsi utama chatbot bagi mayoritas populasi adalah untuk pemenuhan kebutuhan kasual dan pengembangan profesional. Karakteristik Cluster 9, yang secara spesifik menonjol pada fitur tips dan resume (Gambar 6), merepresentasikan kelompok "Pencari Solusi Karier". Temuan ini sejalan dengan literasi manajemen informasi yang menyatakan bahwa kelompok dengan tujuan kompetensi spesifik cenderung bergantung pada saluran informasi yang fungsional dan terstruktur.

Di sisi lain, kombinasi TF-IDF dan K-Means berhasil mengisolasi kelompok pengguna dengan minat teknis yang sangat terdefinisi pada Cluster 10 (fitur explain, gravity, dan neural). Pemisahan ini membuktikan kekuatan algoritma dalam mengidentifikasi intent dengan literasi teknis atau akademik tinggi dari pengguna informasi umum seperti pada Cluster 5 (news dan latest) [20]. Pengguna pada Cluster 10 cenderung memandang chatbot sebagai alat pendukung penalaran analitis, mirip dengan profil problem solvers yang memanfaatkan teknologi untuk tugas-tugas kompleks. Sementara itu, Cluster 14 yang berfokus pada fitur suggest dan movies menunjukkan adanya segmen pencari hiburan yang memerlukan pendekatan konten yang lebih naratif dan personal.

5. KESIMPULAN DAN SARAN

Penelitian ini menyimpulkan bahwa algoritma K-Means dengan optimasi K = 54 berhasil mengidentifikasi struktur *intent* percakapan yang sangat bervariasi, dengan lima cluster dominan yang merepresentasikan spektrum kebutuhan utama mulai dari edukasi teknis, pengembangan karier, informasi terkini, hingga rekomendasi gaya hidup. Keberhasilan model dalam memisahkan kelompok dengan literasi teknis tinggi (Cluster 10) dari kelompok informasi umum (Cluster 5) membuktikan bahwa pemrosesan teks menggunakan TF-IDF mampu menghasilkan representasi fitur yang cukup kuat untuk mendukung pengelompokan *intent* yang granular. Sebagai saran untuk penelitian selanjutnya, disarankan untuk menerapkan pemantauan longitudinal guna memahami trajektori perubahan perilaku pengguna antar cluster dari waktu ke waktu, serta mengeksplorasi penggunaan metode *embedding* yang lebih lanjut seperti Word2Vec atau BERT untuk dibandingkan dengan hasil TF-IDF dalam menangkap konteks semantik yang lebih dalam.

DAFTAR PUSTAKA

[1] A. Puspitasari, “Natural Language Processing (NLP) Technology for Chatbot Website,” *J. Tek. Inform. dan Sist. Inf.*, vol. 10, no. 1, pp. 15–24, 2024.

[2] Setiawan, R., “Comparative Analysis of Clustering Approaches in Assessing ChatGPT User Behavior,” *J. Media Inform. Budidarma*, vol. 8, no. 2, pp. 550–558, 2024.

[3] Jemimah, S., “Intent detection in AI chatbots: a comprehensive review of techniques and the role of external knowledge,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 9, no. 1, pp. 112–120, 2025.

[4] I. N. Setiawan, R., Adnyana, “Improving Helpdesk Chatbot Performance with Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity Models,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 4, pp. 789–796, 2023.

[5] R. . Setiawan, A.; Pratama, “Analisis Pengelompokan Topik Keluhan Mahasiswa Berdasarkan Makna Semantik Menggunakan IndoBERT dan K-Means Clustering,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 9, no. 1, pp. 110–118, 2025.

[6] R. C. T. H. Agustine, V.; Ashari, I. F.; Permana, “Clustering of Regions in Lampung Province based on social and economic aspects using the K-Means algorithm with PCA optimization,” *J. Ekon. Stud. Pembang.*, vol. 26, no. 2, pp. 193–201, 2025.

[7] Ardiansyah, B., “Penerapan K-Means Clustering dan Vektorisasi TF-IDF untuk Identifikasi dan Pemetaan Topik Publik Tweet,” *J. Ilm. Sist. Inf.*, vol. 11, no. 1, pp. 45–52, 2025.

[8] I. P. A. E. Purniawan, I. M. A., Sasmita, G. M.

A., Pratama, "Clustering Berita Menggunakan Algoritma TF-IDF dan K-Means dengan Memanfaatkan Sumber Data Crawling pada Situs Detik.com," *JITTER (Jurnal Ilm. Teknol. dan Komputer)*, vol. 3, no. 1, pp. 1–8, 2022.

[9] S. Kholidah and N. Rismawati, "Perbandingan Penerapan Text Mining Pada Aplikasi Wargaku Surabaya Untuk Aduan Masyarakat Terhadap Satpol Pp Kota Surabaya Dengan Algoritma K-Means Clustering Dan Fuzzy C-Means Clustering," *J. Ilm. Teknol. Inf. Asia*, vol. 19, no. 1, pp. 34–42, 2025.

[10] J. Kandaraj, R. Kannan, and F. Andres, "Intent detection in AI chatbots: a comprehensive review of techniques and the role of external knowledge," *IAES Int. J. Artif. Intell.*, vol. 14, no. 5, pp. 4250–4259, 2025.

[11] G. H. Setiawan and I. M. B. Adnyana, "Improving Helpdesk Chatbot Performance with Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity Models," *JAIC (Journal Appl. Informatics Comput.*, vol. 7, no. 2, pp. 252–257, 2023.

[12] F. Ardiansyah, R. Mupashal, D. Mauludin, M. R. Fachriri, and P. Rosyani, "Penerapan K-Means Clustering dan Vektorisasi TF-IDF untuk Identifikasi dan Pemetaan Topik Publik Tweet Pendidikan UKT, COVID-19, dan Kereta Cepat," *J. AI dan SPK J. Artif. Intell. dan Sist. Penunjang Keputusan*, vol. 3, no. 2, pp. 143–150, 2025.

[13] A. Singh and A. S. Namin, "A survey on chatbots and large language models: Testing and evaluation techniques," *J. Elsevier*, vol. 29, no. 1, pp. 1–18, 2025.

[14] M. I. Mansiz and Z. Fatah, "Pengelompokan Pengguna Media Sosial Berdasarkan Pola Interaksi Menggunakan K-Means," *Gudang J. Multidisiplin Ilmu*, vol. 2, no. 11, pp. 388–397, 2024.

[15] M. A. Haq, W. Purnomo, and N. Y. Setiawan, "Analisis Clustering Topik Survey menggunakan Algoritme K-Means (Studi Kasus: Kudata)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 7, pp. 3498–3506, 2023.

[16] R. Rasyid, Y. Salim, and R. Ramdaniah, "Sistem Informasi Pemetaan Kebutuhan Tenaga Kerja Guru Berbasis Web Menggunakan Metode K-Means," *Bul. Sist. Inf. dan Teknol. Islam*, vol. 4, no. 1, pp. 1–9, 2023.

[17] D. Setiawan, D. Arsa, L. E. Fitri, and F. F. P. Zahardy, "Comparative Analysis of Clustering Approaches in Assessing ChatGPT User Behavior," *J. Media Inform. Budidarma*, vol. 8, no. 2, pp. 550–558, 2024.

[18] R. Suryodiningrat and A. Pratama, "K-Means Clustering for Segmenting AI Survey Respondents: Analysis of Information Sources and Impact Perceptions," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 9, no. 2, pp. 210–218, 2025.

[19] A. Hidayatullah and R. Nuraini, "Uncovering Key Topics in Indonesian Political Discourse Through Twitter Analysis Using Clustering Methods," *J. Edukasi dan Penelit. Inform.*, vol. 11, no. 1, pp. 55–63, 2025.

[20] A. Sharafi, "Identifikasi Pola Diskusi Publik mengenai Pemindahan Ibu Kota Negara Menggunakan Analisis TF-IDF dan K-Means," *J. Media Inform. Budidarma*, vol. 8, no. 3, pp. 1021–1030, 2024.