Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.4, August 2025, pp. 1715~1721
e-ISSN: 2548-6861

Topic Clustering of Student Complaints Based on Semantic Meaning Using the indoBERT and K-Means Models

Gede Herdian Setiawan [1]*, Made Doddy Adi Pranata [2], Ida Bagus Alit Arimbawa [3]*, I Wayan Paramarta Giri [4]*, Ni Putu Leona Carisa Dayani [5]****

* Sistem Komputer, Institut Teknologi dan Bisnis STIKOM Bali
** Bisnis Digital, Institu Teknologi dan Bisnis STIKOM Bali
herdian@stikom-bali.ac.id [1], doddy@stikom-bali.ac.id [2], 230010061@stikom-bali.ac.id [3], 230010104@stikom-bali.ac.id [4], 230010094@stikom-bali.ac.id [5]

Article Info

Article history:

Received 2025-07-05
Revised 2025-07-16
Accepted 2025-07-19

Keyword:

NLP,
IndoBERT,
K-Means,
Clustering.

ABSTRACT

This study applies Natural Language Processing (NLP) technology to extract and cluster information from student complaint text data. The model used is IndoBERT, a variant of BERT (Bidirectional Encoder Representations from Transformers) that has been adapted for the Indonesian language. The main objective of this research is to perform topic clustering based on semantic similarity. The process begins with data collection and cleaning, followed by tokenization and text normalization. Each complaint is transformed into a vector representation through IndoBERT embeddings, which are then used as input for the K-Means clustering algorithm. Evaluation is conducted using various metrics, and the results of the Silhouette Score and Elbow Method indicate that the optimal number of clusters is four. Cluster visualization using the t-distributed Stochastic Neighbor Embedding (t-SNE) method reinforces these findings by displaying four fairly distinct groups of complaints, although one cluster appears dispersed and less well-defined, indicating possible topic overlap. The quality of topics within each cluster is evaluated using the Topic Coherence (c_v) metric, where Cluster 3 achieved the highest score of 0.7084. The topics in this cluster highlight critical issues such as campus facilities, lecturer quality, and information delivery systems. Overall, the four resulting clusters reflect central themes: Facilities, Expectations or Impressions, Services, and Academic Lectures. These results are expected to serve as a reference for institutions in formulating service improvement policies based on student complaint analysis.

I. PENDAHULUAN

Perguruan tinggi sebagai institusi pendidikan tinggi memiliki peran penting dalam membentuk generasi muda. Salah satu aspek penting dalam pengelolaan perguruan tinggi adalah pengelolaan aduan mahasiswa. Aduan yang diajukan mahasiswa mencerminkan berbagai permasalahan yang mereka hadapi, mulai dari masalah akademik, fasilitas, hingga pelayanan. Namun, jumlah aduan yang terus meningkat membuat pengelolaan aduan menjadi semakin kompleks. Untuk mengatasi hal ini, diperlukan sebuah sistem yang dapat mengidentifikasi topik aduan berdasarkan topiknya. Dengan demikian, pihak perguruan tinggi dapat lebih efektif dalam merespon dan menyelesaikan aduan mahasiswa [1]. Teknologi pemrosesan bahasa alami *(Natural Language Processing atau NLP)* telah berkembang pesat dan menawarkan solusi dalam mengekstraksi informasi dari teks, termasuk klasterisasi topik. Salah satu model NLP yang dapat digunakan adalah BERT *(Bidirectional Encoder Representations from Transformers)*.

IndoBERT, sebagai versi BERT yang dilatih khusus untuk Bahasa Indonesia, memiliki kemampuan untuk menangkap makna semantik dari kalimat secara lebih mendalam. Dengan menggunakan embedding dari IndoBERT, aduan mahasiswa dapat diubah menjadi representasi vektor yang mengandung informasi kontekstual [2], [3]. Untuk mengelompokkan aduan berdasarkan kemiripan semantisnya, menggunakan metode

K-Means clustering. K-Means mampu membagi data teks ke dalam kluster-kluster berdasarkan jarak vektor dari hasil embedding IndoBERT.

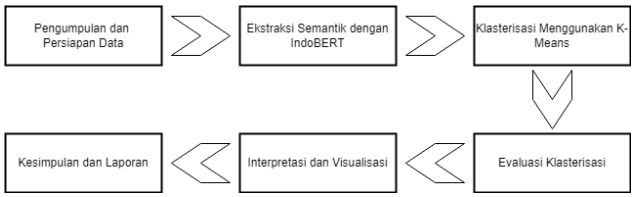
Dengan demikian, klasterisasi ini memungkinkan untuk mengidentifikasi topik aduan yang paling sering muncul, seperti keluhan tentang kualitas pengajaran, keterlambatan layanan administrasi, atau fasilitas kampus yang kurang memadai. Penelitian sebelumnya mengenai klasterisasi topik pada dokumen teks telah menggunakan berbagai metode seperti menerapkan model LDA (*Latent Dirichlet Allocation*)[1], [4]. LDA cenderung mengasumsikan bahwa setiap dokumen hanya membahas satu topik utama, tanpa melihat cakupan beberapa topik sekaligus. Selain itu, LDA kurang mampu menangkap konteks kata dalam kalimat. Model clustering seperti K-Means, DBSCAN *(Density-Based Spatial Clustering of Applications with Noise)* hanya melihat kemiripan vektor fitur, tanpa mempertimbangkan konteks kata dalam kalimat. Akibatnya, kata-kata yang memiliki makna berbeda dalam konteks yang berbeda dapat dikelompokkan dalam satu cluster [5], [6], [7].

Metode clustering dengan menggunakan kombinasi model K-Means dan Word Ebedding untuk ekstraksi fitur menunjukkan hasil baik dalam mengenali makna pada teks namun model word embedding mempengaruhi hasil klasterisasi berita Bahasa Indonesia. Penggunaan model bahasa umum seringkali menghasilkan hasil yang kurang optimal untuk bahasa Indonesia. Selain itu, sebagian besar penelitian hanya mengandalkan algoritma K-Means untuk *clustering*, tanpa mempertimbangkan algoritma *clustering* lainnya yang mungkin lebih cocok untuk data teks [8], [9], [10]. Penelitian ini bertujuan untuk mengatasi keterbatasan-keterbatasan tersebut dengan menggunakan model IndoBERT yang lebih sesuai untuk bahasa Indonesia, dan melakukan evaluasi yang lebih komprehensif terhadap kinerja model [11], [12], [13]. Penggunaan IndoBERT dalam penelitian ini diharapkan dapat memberikan keunggulan karena model ini telah dilatih secara khusus dengan data Bahasa Indonesia, sehingga mampu menangkap nuansa bahasa yang unik di dalam aduan mahasiswa.

Sementara itu, K-Means memberikan fleksibilitas dalam mengelompokkan data dengan skala besar dan kompleks. Kombinasi kedua metode ini tidak hanya membantu dalam mengelompokkan aduan, tetapi juga menyediakan wawasan yang lebih jelas terkait permasalahan utama yang dihadapi oleh mahasiswa, sehingga memungkinkan pengambilan keputusan yang lebih cepat dan akurat. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengelolaan data pengaduan di institusi pendidikan, mempercepat proses respons, dan meningkatkan kepuasan mahasiswa terhadap layanan yang diberikan oleh kampus. Dengan teknologi yang semakin berkembang, otomatisasi dalam klasterisasi topik aduan dapat membantu perguruan tinggi untuk menjadi lebih responsif dan efisien dalam menanggapi kebutuhan mahasiswa.

II. METODE

Metode atau langkah penelitian seperti ditujukan pada Gambar 1 sebagai berikut.



Gambar 1. Langkah Penelitian

A. Pengumpulan dan persiapan data

Pengumpulan data dari data aduan mahasiswa dikumpulkan berupa teks yang diambil dari sistem pengaduan. Data aduan dijadukan sebagai sumber data untuk selanjutnya dilakukan proses Preprocessing Data sebagai langkah untuk pembersihan teks, seperti, normalisasi kata (mengubah huruf kapital menjadi huruf kecil, menghilangkan tanda baca), dan tokenisasi teks.

B. Ekstraksi semantik dengan indoBERT

Penerapan IndoBERT, model berbasis transformer yang sudah dilatih pada bahasa Indonesia, digunakan untuk menangkap representasi semantik dari setiap aduan. IndoBERT akan menghasilkan vektor embedding yang merepresentasikan makna setiap aduan. Melakukan Fine-tuning dengan data aduan mahasiswa agar lebih spesifik menangkap makna dalam konteks pendidikan atau lingkungan mahasiswa.

C. Klasterisasi menggunakn K-Means

Setelah mendapatkan vektor embedding dari IndoBERT, metode K-Means diterapkan untuk melakukan klasterisasi. Algoritma ini akan mengelompokkan aduan berdasarkan kemiripan vektor embedding, sehingga aduan dengan makna serupa akan berada dalam satu kelompok.

D. Evaluasi hasil klaster

Evaluasi klasterisasi dilakukan dengan metrik Koherensi seperti Coherency Score untuk mengukur sejauh mana kata-kata dalam sebuah klaster memiliki hubungan semantik yang kuat dan melihat secara manual hasil ekstraksi kata dalam sebuah klaster akan muncul bersama-sama dalam sebuah dokumen.

E. Interpertasi dan Visualisasi

Melakukan analisis terhadap masing-masing klaster untuk memahami topik utama dari aduan yang terkumpul di dalamnya dan melakukan analisis manual terhadap klaster menggunakan visualisasi TSNE (t-distributed Stochastic Neighbor Embedding)

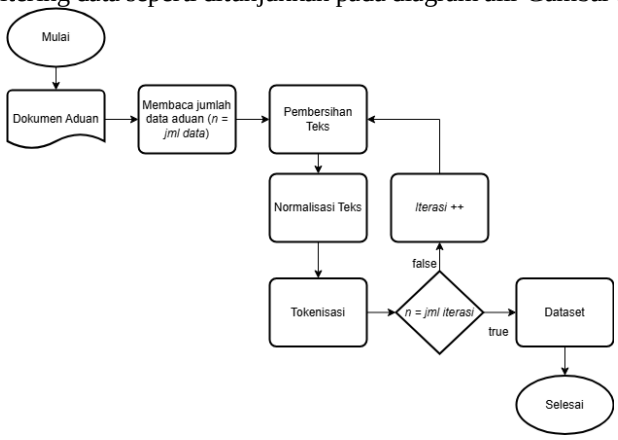
F. Kesimpulan dan laporan

Membuat kesimpulan hasil penelitian, membuat laporan dan melakukan publikasi ilmiah.

III. HASIL DAN PEMBAHASAN

A. Pengumpulan dan persiapan data

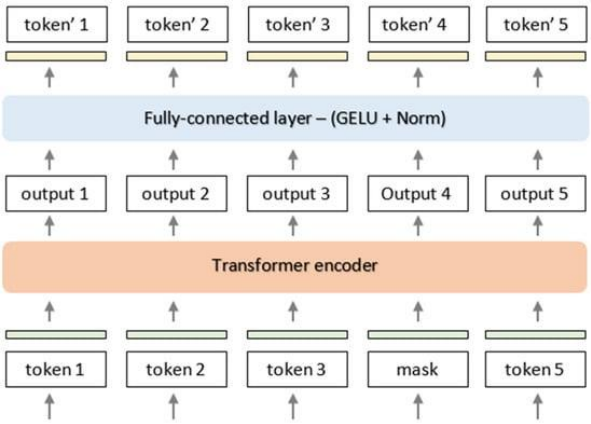
Data dikumpulkan melalui basis data pengaduan Mahasiswa pada sistem informasi institusi dengan rentan waktu Maret 2024 sampai dengan Pebruari 2025, pada tahap ini jumlah data dikumpulkan 5.000 data pengaduan mahasiswa, data yang dikumpulkan disajikan pada link berikut : https://github.com/herdianset/clustering-indoBERT.git. Data pengaduan selanjutnya dilakukan proses filtering data seperti ditunjukkan pada diagram alir Gambar 2.



Gambar 2. Pra Pemrosesan Data

Gambar 2 menyajikan proses persiapan dataset. Dokumen aduan mahasiswa yang terkumpul berupa teks aduan yang menjadi sumber utama. Setelah data terkumpul, dilakukan pembersihan teks untuk menghilangkan noise seperti karakter khusus, spasi berlebih, dan tanda baca yang tidak relevan. Selanjutnya, dilakukan normalisasi teks dengan mengubah seluruh teks menjadi huruf kecil untuk menjaga konsistensi. Tahap terakhir adalah tokenisasi, yaitu memecah teks menjadi unit-unit kecil yang disebut token, sehingga dapat diolah lebih lanjut oleh model. Proses ini memastikan data yang digunakan bersih dan siap untuk dianalisis menggunakan model IndoBERT.

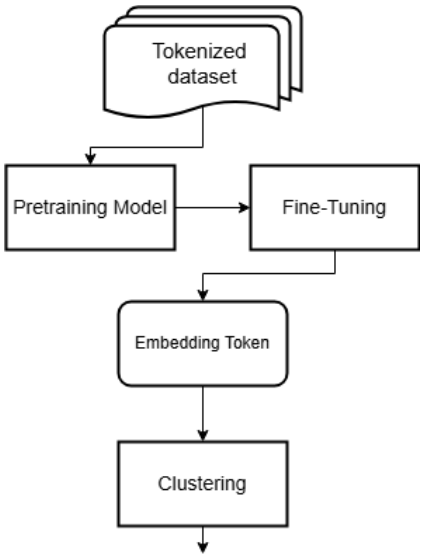
B. Ekstraksi semantic dengan model indoBERT



Gambar 3. Arsitektur BERT

IndoBERT adalah model berbasis BERT yang dilatih khusus untuk bahasa Indonesia, menggunakan arsitektur Transformer untuk menangkap konteks kata dalam teks secara bidirectional. Model ini terdiri dari beberapa lapisan encoder yang menerapkan mekanisme self-attention, memungkinkan model untuk memahami hubungan antar kata dalam kalimat. IndoBERT dilatih dengan dua tugas utama, yaitu Masked Language Modeling (MLM) dan Next Sentence Prediction (NSP). Dengan tokenizer berbasis WordPiece, IndoBERT dapat menangani kata atau subkata yang tidak ada dalam kosakata pelatihan, menjadikannya efektif untuk teks yang bervariasi [14], [15], [16]. Secara umum arsitektur model BERT ditujukan pada Gambar 3.

Pada penelitian ini proses dimulai dengan tokenisasi *dataset*, di mana teks mentah diubah menjadi token atau sub-kata menggunakan tokenizer berbasis WordPiece. Tokenisasi ini memungkinkan model untuk menangani kata-kata yang tidak ada dalam kosakata pelatihan dengan membagi kata menjadi unit yang lebih kecil. Selanjutnya, pada tahap *pretraining*, IndoBERT dilatih menggunakan dua tugas utama, yaitu *Masked Language Modeling* (MLM) dan *Next Sentence Prediction* (NSP). Sebagian kata dalam kalimat dihilangkan secara acak, dan model dipaksa menebak kata yang hilang berdasarkan konteks. Sementara NSP bertujuan untuk memprediksi apakah dua kalimat saling berhubungan. *Pretraining* ini dilakukan dengan dataset besar bahasa Indonesia untuk membangun pemahaman semantik. Setelah itu, IndoBERT dapat fine-tuned pada *dataset* aduan Mahasiswa dengan tujuan untuk memahami konteks dalam lingkup yang lebih spesifik. Dilanjutkan dengan proses *embedding* token, token yang dihasilkan dari tokenisasi diubah menjadi representasi vektor melalui token *embeddings* dan *positional embeddings*, yang memungkinkan model untuk memahami makna dan urutan kata dalam kalimat untuk selanjutnya dilakukan proses Clustering. Proses ini ditujukan pada Gambar 4.



Gambar 5. Eksraksi Semantik dengan Model indoBERT

Proses *Fine-tuning* parameter yang digunakan disajikan pada Tabel 1.

TABEL 1
PARAMETER FINE-TUNING MODEL

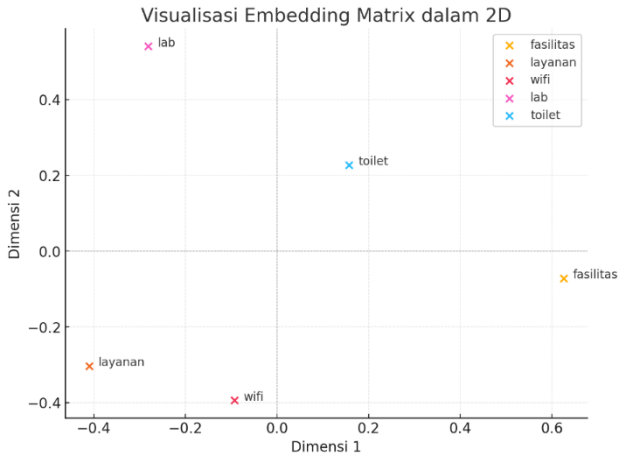
No	Parameter	Nilai
1	*Epoch*	3
2	*Batch Size*	32
3	*Learning Rate*	5e-5
4	*Max length*	512

Nilai pada parameter yang disajikan pada Tabel 1 masih perlu penyesuaian sehingga mendapat hasil yang maksimal. Setelah proses *fine-tuning* selesai selanjutnya membentuk *embeding matrix. Embeding matrix* merupakan komponen inti yang menyimpan representasi vektor dari semua token dalam kosakata model. Matriks ini memungkinkan model untuk memahami dan memproses kata-kata dalam bentuk numerik, dengan mempertimbangkan makna dan hubungan semantik antar kata. Dimensi embedding indbert-base-p1 yang digunakan pada penelitian ini adalah 768 Tabel 2 menyajikan simulasi *embeding matrix* untuk kosakata : *["fasilitas", "layanan", 'wifi","lab","toilet"]*

TABEL 2
EMBEDING MATRIX

Kata	e_1	e_2	e_3	e_4
Fasilitas	0.3745	0.9507	0.7320	0.5987
Layanan	0.1560	0.1560	0.0581	0.8662
Wifi	0.6011	0.7081	0.0206	0.9699
lab	0.8324	0.2123	0.1818	0.1834
toilet	0.3042	0.5248	0.4319	0.2912

Setiap kata direpresantisakan sebagai baris dalam matriks embedding, contoh pada kata fasilitas memiliki *embeding* : *[0.3745, 0.9507, 0.7320, 0.5987]*. hal ini memungkinkan memahami semantik antar kata, kata yang sering muncul bersama memiliki embedding yang lebih dekat pada vektor. *Embedding* dalam visualisasi 2D ditunjukkan pada Gambar 5.



Gambar 5 Visualisasi Embedding Matrix

Gambar 5 menunjukkan visualisasi embedding matrix dari kosakata *["fasilitas", "layanan", 'wifi","lab","toilet"]* dalam 2D, setiap titik mewakili *embedding* dari sebuah kata, kata yang lebih dekat satu sama lain memiliki hubungan semantik kuat. Pada penelitian ini hasil *embedding matrix* disajikan pada Gambar 6.

```
[[-0.23734617 -0.18003097 -0.4773155  ... -0.271559   -0.72630495  0.9272196 ]
 [ 0.5872368  -0.04550786 -0.32442334 ... -0.71305066 -1.3149874  -0.09575679]
 [ 0.21831028 -0.4487409  -0.37828344 ... -0.21616961  0.05635373  0.76014465]
 ...
 [ 0.3555373  -0.40279084 -0.08733421 ... -0.40522502 -0.5455914  -0.23602232]
 [-0.15544419 -0.21479344 -0.20598498 ... -0.5954086  -0.8321377   0.19124006]
 [-0.36077428  0.17542565 -0.8119596  ... -0.24585095 -0.9217965   0.6350337 ]]
```

Gambar 6. Hasil Embedding Matrix

C. Sub-Bab

Setelah embedding matrix terbentuk dengan model indoBERT selanjutnya melakukan klustering terhadap topik pada pengaduan, sebelum melakukan proses clustering terlebih dahulu dilakukan reduksi embedding matrix dengan tujuan klusterisasi menjadi lebih cepat karena berdimensi lebih rendah, mengurangi noise / fitur yang tidak relevan atau redundan dapat memperkenalkan noise ke algoritma klasterisasi, menyebabkan pembentukan klaster yang kurang optimal. dan meningkatkan separasi antar klaster. Pada penelitian ini untuk mereduksi embedding matrix menggunakan model PCA *(Principal Component Analysis)* dengan langkah sebagai berikut :

Menghitung matrix Kovarians :

$$C = \frac{1}{n-1} e^T X$$

Dimana

C = Kovaronus,

X = embedding matrix,

V_k *= matriks eigenvectors*

Setelah mereduksi embedding matrix, selanjutnya penentuan jumlah klaster yang paling ideal, proses ini menggunakan algoritma *Silhouette Score* sebagai berikut :

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Dimana :

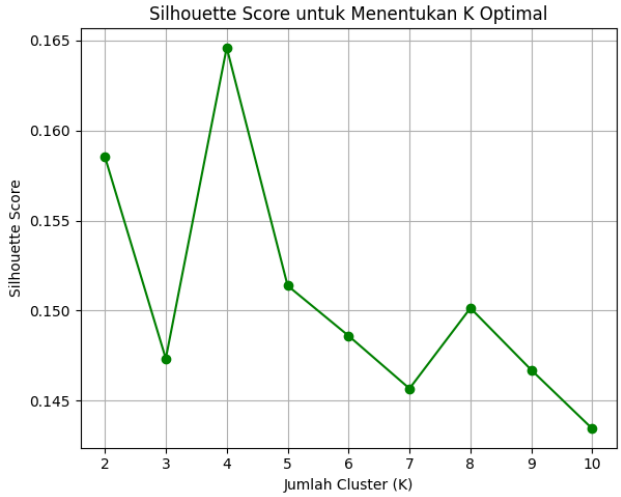
a(i) = rata-rata jarak data *i* ke anggota klaster

b(i) = rata-rata jarak data *i* ke klaster terdekat

S(i) = klaster mendekati 1 (baik) mendekati 0 (batas klaster) -1 (klaster buruk)

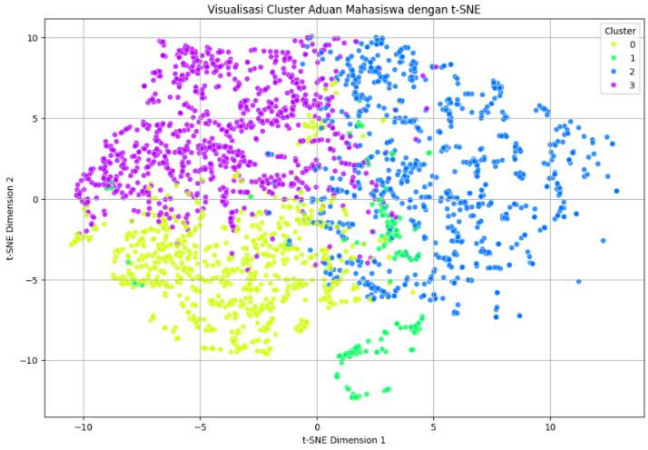
Untuk menentukan jumlah kluster *k* yang paling ideal adalah nilai *Silhouette Score* paling tinggi [14].

Pada penelitian ini jumlah kluster yang paling ideal adalah 4 kluster dengan score 0.1646, hasil perhitungan *Silhouette Score* disajikan dalam Gambar 7. Jumlah iterasi pada K-Means adalah 300, radom seed adalah 48 dan n_init = "auto" yang akan ditentukan nilai terbaik oleh pustaka Scikit-learn.



Gambar 7 Nilai *Silhouette Score*

Visualisasi klaster dengan t-SNE disajikan pada Gambar 8.



Gambar 8 Visualisai Cluster

Gambar 8 menunjukkan visualisasi klaster menggunakan t-SNE, hasil klasterisasi data ke dalam empat klaster sesuai dengan *Silhouette Score* tertinggi (terbaik), masing-masing kluster ditandai dengan warna kuning (klaster 0), hijau (klater 1), biru (klaster 2) dan ungu (klaster 3). Klaster kuning terlihat sebagai cluster yang cukup padat dan terkonsentrasi di bagian kiri bawah grafik. Klaster hijau merupakan klaster dengan jumlah anggota yang paling sedikit. Titik-titiknya tampak lebih tersebar dan berada di antara klaster 0 dan klaster 2, serta beberapa di bagian bawah tengah. Klaster biru merupakan salah satu klaster terbesar dan paling dominan, menempati sebagian besar sisi kanan grafik. Titik-titiknya tampak menyebar cukup luas. Klaster ungu juga merupakan cluster yang besar dan terkonsentrasi di bagian kiri atas grafik, menunjukkan kepadatan yang cukup tinggi.

Sparasi antar klaster Kuning dan Ungu menunjukkan separasi yang relatif baik satu sama lain dan juga terhadap kluster Biru. Meskipun ada beberapa titik yang mungkin berdekatan di perbatasan, secara umum mereka membentuk kelompok yang berbeda. Klaster Hijau tampak kurang terdefinisi dengan baik dan tersebar di antara kluster-kluster lain, terutama dengan klaster kuning dan klaster biru. Hal ini bisa mengindikasikan bahwa aduan dalam klaster Hijau mungkin memiliki karakteristik yang tumpang tindih atau merupakan kasus-kasus yang lebih ambigu yang tidak secara jelas masuk ke dalam klaster. Secara umum berdasarkan Visualisasi Klaster dapat menggambarkan bagaimana aduan mahasiswa dapat dikelompokkan berdasarkan kesamaan karakteristiknya.

D. Evaluasi hasil klaster berdasarkan kualitas topik secara semantik

Untuk menjamin bahwa topik dalam setiap klaster memiliki keterkaitan makna yang baik, dilakukan evaluasi menggunakan metode *Topic Coherence*. Metode ini secara khusus dirancang untuk menilai sejauh mana kumpulan kata dalam suatu klaster (topik) membentuk suatu konsep yang bermakna secara semantik dan mudah dipahami. Dalam penelitian ini digunakan metrik koherensi *c_v*, yang mengombinasikan *Normalized Pointwise Mutual Information* (NPMI), kesamaan kosinus dari vektor kata (seperti Word2Vec atau GloVe), serta pendekatan *sliding window* untuk menghitung frekuensi kemunculan kata-kata utama dalam topik secara bersamaan [17]. Nilai koherensi untuk setiap klaster ditampilkan pada Tabel 3.

TABEL 3.
NILAI KOHERENSI SETIAP KLASTER

No	*Cluster*	*c_v Schore*
1	*Cluster 0*	0.4059
2	*Cluster 1*	0.4023
3	*Cluster 2*	0.3677
4	*Cluster 3*	0.7084

Sebagaimana disajikan pada Tabel 3, klaster 3 menunjukkan ringkasan topik yang paling representatif, ditandai dengan skor *c_v* yang cukup tinggi, yaitu 0.7084. Skor *c_v* yang tinggi ini mengindikasikan bahwa kata-kata dalam topik tersebut sering muncul secara bersamaan dalam dokumen aduan. Topik utama yang diangkat dalam klaster ini berkaitan dengan saran untuk perbaikan di beberapa aspek penting di lingkungan kampus, seperti fasilitas, mutu dosen, dan sistem penyampaian informasi. Klaster 0 memberikan penjabaran lebih rinci dari salah satu subtopik dalam Klaster 3, yaitu terkait fasilitas, dengan memberikan bukti konkret mengenai aspek fasilitas mana saja yang memerlukan perhatian. Sementara itu, klaster 1 dan klaster 2 tergolong sebagai topik yang lemah atau saling tumpang tindih. Keduanya cenderung menggambarkan sentimen positif yang umum, namun kurang jelas dan memiliki kemiripan isi, sebagaimana tergambarkan dalam visualisasi pada Gambar 8.

E. Visualisasi kata aduan menggunakan Word Cloud

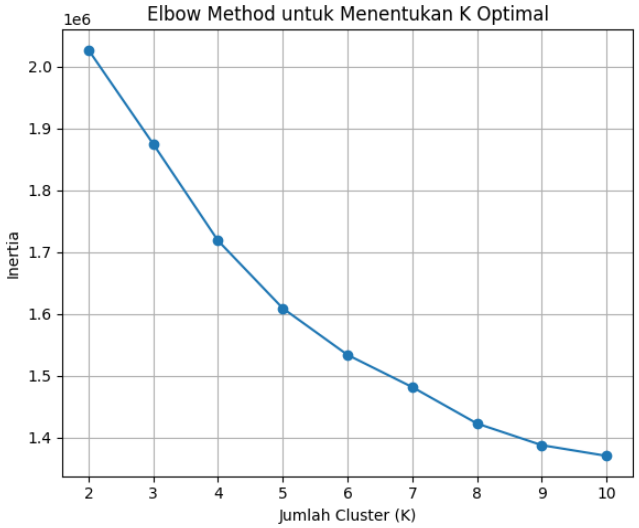
Ekstraksi dan visualisasi kata pada kalimat aduan menggunakan Word Cloud bertujuan untuk memberikan gambaran visual yang mudah dipahami mengenai topik utama dari setiap klaster aduan. Word Cloud menampilkan kata-kata

yang paling sering muncul dalam setiap kelompok data, di mana ukuran kata mencerminkan frekuensi kemunculannya. Word cloud pada aduan mahasiswa ditunjukkan pada Gambar 9.

Gambar 9. Word Cloud Topik Aduan Mahasiswa

Seperti yang ditujukan pada gambar 9. Kata *"fasilitas"* menjadi yang paling menonjol, menunjukkan bahwa sebagian besar aduan mahasiswa terkait dengan kondisi atau ketersediaan fasilitas kampus. Kata *"harap"* dan *"kampus"* mengindikasikan adanya harapan mahasiswa terhadap perbaikan lingkungan kampus secara umum. Sementara itu, kata *"kelas"* dan *"ajar"* mencerminkan isu-isu yang berkaitan dengan kegiatan belajar mengajar, seperti kualitas ruang kelas, proses pembelajaran, atau metode pengajaran dosen. Selain itu, terdapat kata-kata pendukung seperti *"toilet", "parkir", "wifi", "layanan",* dan *"teknologi"* yang memperkuat bahwa aspek fisik dan layanan penunjang akademik menjadi perhatian utama mahasiswa. Kata *"dosen", "program", dan "tugas"* juga menunjukkan adanya perhatian terhadap aspek akademik dan pengajaran. Berdasarkan visualisasi dari world cloud sesuai dengan topik klaster pada ke-empat klaster yaitu fasilitas, kesan dan harapan umum, pelayanan dan proses perkuliahan.

Untuk validasi jumlah klaster, digunakan visualisasi Elbow Method seperti terlihat pada Gambar 10.



Gambar 10. Jumlah Cluster Optimal dengan Elbow Method

Elbow method pada Gambar 10 menunjukkan garis Elbow (siku) berada di sekitar K=4 atau K=5. Titik K=4 menunjukkan perubahan laju penurunan yang cukup jelas, dan K=5 juga masih menunjukkan perbaikan sebelum kurva menjadi lebih datar. Penelitian ini telah menggunakan jumlah klaster yang paling optimal yaitu empat klaster.

IV. KESIMPULAN

Penelitian ini berhasil mengelompokkan aduan mahasiswa ke dalam beberapa topik utama dengan pendekatan berbasis *semantic clustering*. Proses dimulai dari tahapan praproses data hingga pembentukan representasi semantik menggunakan embedding model IndoBERT. Hasil pengelompokan dengan algoritma K-Means menunjukkan bahwa jumlah klaster optimal adalah empat, sebagaimana ditunjukkan oleh nilai *Silhouette Score* tertinggi dan didukung oleh analisis visual menggunakan *Elbow Method*. Visualisasi dua dimensi melalui *t-SNE* juga memperlihatkan struktur klaster yang cukup terpisah, meskipun terdapat satu klaster yang menunjukkan indikasi tumpang tindih.

Evaluasi lebih lanjut terhadap kualitas topik dilakukan menggunakan metrik *Topic Coherence* dengan pendekatan *c_v*, yang menggabungkan NPMI, kemiripan kosinus, dan sliding window. Klaster 3 memperoleh skor *c_v* tertinggi sebesar 0.7084, menandakan bahwa topik dalam klaster ini memiliki kohesi semantik yang kuat. Topik dalam klaster tersebut mencerminkan isu-isu strategis seperti perbaikan fasilitas kampus, kualitas dosen, dan sistem informasi. Sementara itu, Klaster 0 mengelaborasi lebih lanjut topik fasilitas, dan Klaster 1 serta Klaster 2 dinilai lemah karena mengandung sentimen positif yang umum namun tidak memiliki fokus topik yang jelas, serta menunjukkan potensi tumpang tindih.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa aduan mahasiswa dapat dikelompokkan ke dalam empat tema utama yang bermakna secara semantik, yaitu: Fasilitas, Harapan atau Kesan, Layanan, dan Perkuliahan. Pendekatan ini tidak hanya membantu memahami pola aduan, tetapi juga menyediakan dasar yang kuat untuk perumusan kebijakan peningkatan kualitas layanan di lingkungan kampus. Penelitian selanjutnya berfokus pada analisis mendalam terhadap klaster ambigu, penajaman topik setiap klaster, serta optimalisasi metodologi melalui model embedding lain dan menggunakan metode klastering yang berbeda.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih kepada Institut Teknologi dan Bisnis STIKOM Bali yang telah memberikan pembiayaan penelitian ini dan telah mendukung selama berjalannya kegiatan penelitian.

DAFTAR PUSTAKA

[1] G. H. Setiawan, I. Made, B. Adnyana, G. Rai, A. Sugiartha, and K. Budiarta, "Ekstraksi Topik Pada Aduan Mahasiswa Dengan Pendeketan Model Latent Dirichlet Allocation (LDA)," 2024.

[2] J. Huang and X. Zhu, "Deep Semantic Clustering by Partition Confidence Maximisation," 2020.

[3] L. Geni, E. Yulianti, and D. I. Sensuse, "Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 9, no. 3, pp. 746–757, 2023, doi: 10.26555/jiteki.v9i3.26490.

[4] D. A. Rahman, R. B. Waskitho, M. Fajrul, A. U. Nuha, and N. A. Rakhmawati, "Klasterisasi Topik Konten Channel Youtube Gaming Indonesia Menggunakan Latent Dirichlet Allocation," 2021.

[5] H. Tommy Argo Simanjuntak, P. Ephraim Prabowo Silaban, J. Koko Sarasi Manurung, and V. Handayani Sormin, "Klasterisasi Berita Bahasa Indonesia Dengan Menggunakan K-Means Dan Word Embedding," vol. 10, no. 3, pp. 641–652, 2023, doi: 10.25126/jtiik.2023106468.

[6] Z. Vladimir, D. Alamsyah, and W. Widhiarso, "Klasterisasi Topik Skripsi Informatika Dengan Metode DBSCAN," *Jurnal Algoritme*, vol. 3, no. 1, 2022.

[7] R. Siringoringo, "Text Mining dan Klasterisasi Sentimen Pada Ulasan Produk Toko Online," 2019.

[8] M. Riduwan, C. Fatichah, and A. Yuniarti, "Klasterisasi Dokumen Menggunakan Weighted K-Means Berdasarkan Relevansi Topik," 2019.

[9] S. W. Harjono *et al.*, "Klasterisasi Tingkat Penjualan pada Startup Panak.id dengan Algoritma K-Means," *Jurnal Ilmiah Teknologi Informasi Asia*, vol. 17, no. 1, 2023.

[10] M. R. Arief, D. O. Siahaan, and I. Arieshanti, "Klasterisasi Teks Menggunakan Metode Max-Max Roughness (Mmr) Dengan Pengayaan Similaritas Kata," 2010.

[11] S. M. Isa, G. Nico, and M. Permana, "Indobert For Indonesian Fake News Detection," *ICIC Express Letters*, vol. 16, no. 3, pp. 289–297, Mar. 2022, doi: 10.24507/icicel.16.03.289.

[12] S. Saadah, Kaenova Mahendra Auditama, Ananda Affan Fattahila, Fendi Irfan Amorokhman, Annisa Aditsania, and Aniq Atiqi Rohmawati, "Implementation of BERT, IndoBERT, and CNN-LSTM in Classifying Public Opinion about COVID-19 Vaccine in Indonesia," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 4, pp. 648–655, Aug. 2022, doi: 10.29207/resti.v6i4.4215.

[13] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, "Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN," *ILKOM Jurnal Ilmiah*, vol. 14, no. 3, pp. 348–354, Dec. 2022, doi: 10.33096/ilkom.v14i3.1505.348-354.

[14] A. Punhani, N. Faujdar, K. K. Mishra, and M. Subramanian, "Binning-Based Silhouette Approach to Find the Optimal Cluster Using K-Means," *IEEE Access*, vol. 10, pp. 115025–115032, 2022, doi: 10.1109/ACCESS.2022.3215568.

[15] C. H. Miranda, G. Sanchez-Torres, and D. Salcedo, "Exploring the Evolution of Sentiment in Spanish Pandemic Tweets: A Data Analysis Based on a Fine-Tuned BERT Architecture," *Data (Basel)*, vol. 8, no. 6, Jun. 2023, doi: 10.3390/data8060096.

[16] L. Geni, E. Yulianti, and D. I. Sensuse, "Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 9, no. 3, pp. 746–757, 2023, doi: 10.26555/jiteki.v9i3.26490.

[17] A. Alfajri, D. Richasdy, and M. A. Bijaksana, "Topic Modelling Using Non-Negative Matrix Factorization (NMF) for Telkom University Entry Selection from Instagram Comments," *Journal of Computer System and Informatics (JoSYC)*, vol. 3, no. 4, pp. 485–492, Sep. 2022, doi: 10.47065/josyc.v3i4.2212.