JURNAL INSY PRO Information System and Processing
Volume 9, Nomor 2, November 2024 P–ISSN 2597-419X, E-ISSN 2579-468X
Ojs : http://journal.uinalauddin.ac.id/index.php/insypro
Email : insypro@uin-alauddin.ac.id

PEMODELAN TOPIK SARAN MAHASISWA PADA SIMAK UNISMUH MENGGUNAKAN BERTOPIC

Topic Modelling of Student Suggestions on SIMAK UNISMUH Using BERTopic

Fahrim Irhamna Rachman[1)], Rizki Yusliana Bakti[2)] , Titin Wahyuni[3)], Lis Indriani[4)]

[1,2,3,4]Program Studi Informatika Universitas Muhammadiyah Makassar
[1,2,3,4]Jl. Sultan Alauddin No 259, Kota Makassar, Sulawesi Selatan 90123

E-mail: fachrim141020@unismuh.ac.id[1)], Rizkiyusliana@unismuh.ac.id[2)], Titinwahyuni@unismuh.ac.id[3)], 105841108020@student.unismuh.ac.id[4)]

Abstrak – Penelitian ini mengkaji penggunaan algoritma BERTopic untuk pemodelan topik pada data saran mahasiswa yang dikumpulkan melalui SIMAK UNISMUH. Penelitian ini bertujuan untuk mengidentifikasi dan memvisualisasikan pola tematik dalam saran mahasiswa, dengan tujuan meningkatkan layanan akademik dan fasilitas kampus. Studi ini memanfaatkan teknik *Natural Language Processing* (NLP), khususnya BERTopic, yang menggabungkan keunggulan *Bidirectional Encoder Representations from Transformers* (BERT) dengan algoritma *clustering* untuk menghasilkan representasi topik yang kontekstual dan mudah diinterpretasikan. Data penenelitian terdiri dari 232.430 entri saran yang diproses untuk menghilangkan *noise* dan informasi yang tidak relevan, menghasilkan 26.009 entri valid. Entri-entri ini kemudian diproses menggunakan algoritma BERTopic, menghasilkan 9 topik yang berbeda terkait dengan berbagai aspek kehidupan akademik, termasuk metode pengajaran, fasilitas kampus, dan layanan adminstrasi. *Coherence score* sebesar 0.637 menunjukkan konsistensi internal yang kuat dalam topik-topik yang diidentifikasi sementara analisis mengungkapkan area utama di mana universitas dapat meningkatkan layanannya. Temuan dari penelitian ini memberikan wawasan yang dapat ditindaklanjuti bagi administrator universitas, memungkinkan mereka membuat keputusan dan kinerja akademik mahasiswa. Selain itu, penelitian ini berkontribusi pada bidang pemodelan topik dalam konteks pendidikan dan menunujukkan efektivitas BERTopic dalam mengolah data tekstual skala besar.

Kata Kunci: Pemodelan Topik, BERTopic, *Natural Language Processing*

Abstract *– This study examines the use of the BERTopic algorithm for topic modeling on student feedback data collected through the SIMAK UNISMUH. The research aims to identify and visualize thematic patterns in student feedback to improve academic services and campus facilities. This study utilizes Natural Language Processing (NLP) techniques, particularly BERTopic, which combines the advantages of Bidirectional Encoder Representations from Transformers (BERT) with clustering algorithms to produce contextually rich and easily interpretable topic representations. The research data comprises 232,430 feedback entries that were processed to remove noise and irrelevant information, resulting in 26,009 valid entries. These entries were then processed using the BERTopic algorithm, generating nine distinct topics related to various aspects of academic life, including teaching methods, campus facilities, and administrative services. The coherence score of 0.637 indicates strong internal consistency within the identified topics, while the analysis reveals key areas where the university can enhance its services. The findings from this study provide actionable insights for university administrators, enabling them to make informed decisions and improve student academic performance. Additionally, this research contributes to the field of topic modeling in educational contexts and demonstrates the effectiveness of BERTopic in processing large-scale textual data.*

Keywords: *Topic Modelling, BERTopic, Natural Language Processing*

PENDAHULUAN

Perguruan tinggi memiliki tanggung jawab penting dalam menciptakan lingkungan yang mendukung perkembangan akademik, pribadi, dan profesional mahasiswanya. Untuk memenuhi kebutuhan mahasiswa dan meningkatkan kualitas layanan, Universitas Muhammadiyah Makassar (UNISMUH) menggunakan Sistem Informasi Manajemen Akademik (SIMAK), yang memiliki

peran kunci dalam mengelola informasi terkait kegiatan akademik dan administratif. Salah satu fitur dalam SIMAK adalah pengisian saran, yang memungkinkan mahasiswa memberikan umpan balik terkait pengalaman belajar, fasilitas kampus, dan layanan akademik. Mengolah umpan balik ini secara manual bisa menjadi tugas yang memakan waktu dan rawan kesalahan. Oleh karena itu, diperlukan pendekatan yang lebih efisien, seperti teknik pemodelan topik, untuk mengidentifikasi pola tematik dalam data tersebut (Kherwa & Bansal, 2020).

Salah satu pendekatan yang menjanjikan dalam pemodelan topik adalah BERTopic, yang menggabungkan keunggulan model bahasa BERT (*Bidirectional Encoder Repsentations from Transformer*) dengan teknik pemodelan topik. BERT adalah model bahasa yang sangat canggih yang telah terbukti sangat efektif dalam bidang NLP (*Natural Language Processing*) (Lopez-Martinez & Sierra, 2020). Dibandingkan dengan metode populer lainnya seperti LDA (*Lirichlet Dirichlet Allocation*), BERTopic lebih mampu menghasilkan representasi kata yang lebih kaya dan kontekstual dengan mempertimbangkan seluruh konteks kalimat, sementara LDA hanya memperhitungkan kemunculan kata dalam dokumen dan topik (Mustafac, 2021).

Penerapan BERTopic pada umpan balik mahasiswa dari pengisian saran di perguruan tinggi dapat memberi manfaat yang signifikan. Dengan menggunakan teknik ini, kita dapat mengidentifikasi dan mengeksplorasi pola tematik yang muncul dalam umpan balik tersebut secara lebih efisien dan objektif. Ini akan memungkinkan perguruan tinggi untuk mendapatkan wawasan yang lebih dalam tentang preferensi, kebutuhan, dan masalah yang dihadapi oleh mahasiswa mereka.

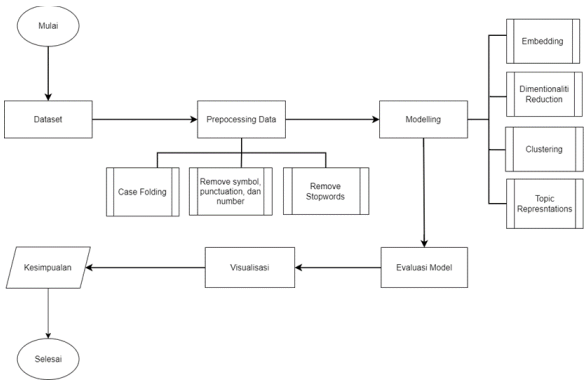
Penelitian ini bertujuan untuk menerapkan teknik pemodelan topik menggunakan BERTopic pada umpan balik mahasiswa dari pengisian saran di Universitas Muhammadiyah Makassar (Unismuh). Dengan demikian, penelitian ini tidak hanya akan memberikan wawasan yang lebih baik tentang pola umpan balik mahasiswa, tetapi juga akan memberikan dasar yang kuat untuk pengambilan keputusan yang lebih baik.

Penelitian ini didasarkan pada referensi dari jurnal penelitian Gibran Giffari Priyatna (2022) dan Herwinsyah (2023) yang membahas pemodelan menggunakan algoritma BERTopic. Perbedaan penelitian ini dengan kedua penelitian tersebut adalah penulis menerapkannya pada data saran mahasiswa dan memvisualisasikan persebaran topiknya.

METODOLOGI PENELITIAN

BERTopic adalah sebuah teknik pemodelan topik yang memanfaatkan kekuatan model bahasa besar seperti BERT untuk menghasilkan topik-topik yang lebih relevan dan mudah diinterpretasi dari suatu kumpulan teks (Lopez-Martinez & Sierra, 2020). Dengan menggabungkan *embedding* kata yang dihasilkan oleh BERT dengan metode *clustering* yang canggih seperti HDBSCAN, BERTopic mampu mengidentifikasi kelompok-kelompok kata yang sering muncul bersama dalam suatu dokumen dan mengelompokkannya menjadi topik-topik yang koheren. Selain itu, BERTopic juga menggunakan pendekatan TF-IDF berbasis kelas untuk memberikan bobot pada kata-kata yang paling representatif untuk setiap topik . Proses ini melibatkan beberapa tahap, mulai dari mengubah kata menjadi representasi numerik (*embedding*), mengurangi dimensi data, melakukan *clustering*, hingga menghitung pentingnya setiap kata dalam setiap klaster. Hasil akhir dari proses ini adalah serangkaian topik yang dapat digunakan untuk memahami isi dari suatu kumpulan dokumen secara lebih mendalam (Atagün *et al*., 2021).

Adapun tahapan dalam perancangan sistem sebagai berikut:

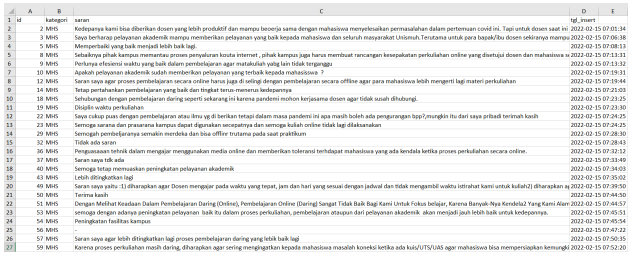


Gambar 1 Perancangan Sistem

Pada gambar 1. Proses dimulai dengan pengambilan Dataset yang diperoleh dari tanggapan pengisian saran mahasiswa pada simak UNISMUH. Jenis data ini dikategorikan sebagai sekunder karena akan terus bertambah seiring berjalannya semester perkuliahan dan masih diberlakukannya angket tersebut. Data yang diperoleh disajikan dalam format

JURNAL INSY PRO Information System and Processing
Volume 9, Nomor 2, November 2024 P–ISSN 2597-419X, E-ISSN 2579-468X
Ojs : http://journal.uinalauddin.ac.id/index.php/insypro
Email : insypro@uin-alauddin.ac.id

CSV dengan total 232.430 baris dan 4 kolom yakni: id, kategori, saran dan tgl_insert. Selanjutnya data yang diperoleh akan disortir dengan hanya mengambil data saran mahasiswa.



	A	B	C	D	E
1	id	kategori	saran	tgl_insert	
2		2 MHS	Kedepanya kami bisa diberikan dosen yang lebih produktif dan mampu beceria sama dengan mahasiswa menyelesaikan permasalahan dalam pertemuan covid ini. Tapi untuk dosen saat ini	2022-02-15 07:01:34	
3		3 MHS	Saya berharap pelayanan akademik mampu memberikan pelayanan yang baik kepada mahasiswa dan seluruh masyarakat Unismuh.Terutama untuk para bapak/ibu dosen sekiranya mampu	2022-02-15 07:06:38	
4		5 MHS	Memperbaiki yang baik menjadi lebih baik lagi.	2022-02-15 07:08:13	
5		8 MHS	Sebaiknya pihak kampus memantau proses penyaluran kouta internet , pihak kampus juga harus membuat rancangan kesepakatan perkuliahan online yang disetujui dosen dan mahasiswa s	2022-02-15 07:13:31	
6		9 MHS	Perlunya efesiensi waktu yang baik dalam pembelajaran agar matakuliah yang lain tidak terganggu	2022-02-15 07:13:32	
7		10 MHS	Apakah pelayanan akademik sudah memberikan pelayanan yang terbaik kepada mahasiswa ?	2022-02-15 07:19:31	
8		12 MHS	Saran saya agar proses pembelajaran secara online harus juga di selingi dengan pembelajaran secara offline agar para mahasiswa lebih mengerti lagi materi perkuliahan	2022-02-15 07:19:44	
9		14 MHS	Tetap pertahankan pembelajaran yang baik dan tingkat terus-menerus kedepannya	2022-02-15 07:21:03	
10		18 MHS	Sehubungan dengan pembelajaran daring seperti sekarang ini karena pandemi mohon kerjasama doen agar tidak susah dihubungi.	2022-02-15 07:23:25	
11		19 MHS	Disiplin waktu perkuliahan	2022-02-15 07:23:30	
12		22 MHS	Saya cukup puas dengan pembelajaran atau ilmu yg di berikan tetapi dalam masa pandemi ini apa masih boleh ada pengurangan bpp?,mungkin itu dari saya pribadi terimah kasih	2022-02-15 07:24:25	
13		23 MHS	Semoga semakin prasarana kampus dapat digunakan secepatnya dan semoga kuliah online tidak lagi dilaksanakan	2022-02-15 07:24:25	
14		29 MHS	Semogah pembelajaranya semakin merdeka dan bisa offline trutama pada saat praktikum	2022-02-15 07:28:30	
15		32 MHS	Tidak ada saran	2022-02-15 07:28:43	
16		36 MHS	Pengusaaaan tehnik dalam mengajar menggunakan media online dan memberikan toleransi terhdapat mahasiswa yang ada kendala ketika proses perkuliahan secara online.	2022-02-15 07:32:12	
17		37 MHS	Saran saya tdk ada	2022-02-15 07:33:49	
18		40 MHS	Semoga tetap memasukan peningkatan pelayanan akademik	2022-02-15 07:34:03	
19		43 MHS	Lebih ditingkatkan lagi	2022-02-15 07:35:02	
20		49 MHS	Saran saya yaitu :1) diharapkan agar Dosen mengajar pada waktu yang tepat, jam dan hari yang sesuai dengan jadwal dan tidak mengambil waktu istirahat kami untuk kuliah2) diharapkan a	2022-02-15 07:39:50	
21		50 MHS	Terima kasih	2022-02-15 07:44:50	
22		51 MHS	Dengan Melihat Keadaan Dalam Pembelajaran Daring (Online), Pembelajaran Online (Daring) Sangat Tidak Baik Bagi Kami Untuk Fokus belajar, Karena Banyak Nya Kendala2 Yang Kami Alam	2022-02-15 07:44:57	
23		53 MHS	semoga dengan adanya peningkatan pelayanan baik itu dalam proses perkuliahan, pembelajaran ataupun dari pelayanan akademik akan menjadi jauh lebih baik untuk kedepannya.	2022-02-15 07:45:51	
24		54 MHS	Peningkatan fasilitas kampus	2022-02-15 07:45:54	
25		56 MHS	-	2022-02-15 07:47:22	
26		57 MHS	Saran saya agar lebih ditingkatkan lagi proses pembelajaran daring yang lebih baik lagi	2022-02-15 07:50:35	
27		59 MHS	Karena proses perkuliahan masih daring, diharapkan agar sering mengingatkan kepada mahasiswa masalah koneksi ketika ada kuis/UTS/UAS agar mahasiswa bisa mempersiapkan kemungki	2022-02-15 07:52:20	

Gambar 2 Dataset

Data dalam format csv akan melalui tahap *prepocessing* sebelum modelling dilakukan. *Prepocessing* pada peneltian ini menggunakan modul *regulareexpression* pada bahasa pemrograman python. Pertama adalah menghilangkan data saran yang memiliki satu kata atau memiliki kurang dari empat huruf. Setelah itu, dilanjutkan dengan proses penyesuaian huruf kecil (*Case Folding*), penghapusan simbol, angka dan tanda baca. Tahap terakhir adalah menghapus kata-kata yang dianggap tidak memberikan kontribusi signifikan dalam analisis, yang dikenal sebagai *stopword removal*.

Apabila Prepocessing data selesai, langkah selanjutnya adalah pemodelan. *Clean* data memasuki tahap prosedur BERTopic. Tahap pertama yang dilakukan adalah *embedding*, yaitu setiap kata dalam ulasan akan diberi bobot sesuai dengan makna semantiknya dalam kalimat. Hasilnya adalah kumpulan kata dalam bentuk matriks berdimensi tinggi yang akan diproses menggunakan algoritma *clustering.* Sebelum proses *clustering* dilakukan, matriks berdimensi tinggi tersebut perlu dikurangi dimensinya menjadi data berdimensi rendah melalui algoritma *dimensionality reduction*. Pada peneltian ini digunakan metode UMAP untuk tujuan tersebut.

Setelah data berhasil direduksi dimensinya, langkah berikutnya adalah melakukan proses *clustering* menggunakan algoritma HDBSCAN. Proses ini akan mengelompokkan kata-kata dalam data menjadi kelompok-kelompok yang merepresentasikan topik-topik yang dibicarakan dalam saran-saran tersebut. Setelah itu, data yang semula berbentuk matriks akan dikembalikan ke bentuk semula dan dilakukan tokenisasi.

Tahap terakhir adalah memberikan bobot pada setiap kata dalam setiap *cluster* menggunakan algoritma c-TF-IDF. Bobot ini akan menunjukkan sejauh mana kata tersebut terkait dengan topik tertentu dan tidak dengan topik lainnya. Hasil ekstraksi topik-topik berdasarkan data akan direpresntasikan dalam bentuk diagram batang (*bartchart*) untuk menampilkan kata-kata yang mewakili setiap topik beserta bobotnya.

Pengujian sistem penelitian ini menggunakan evaluasi model melalui *running time* dan *coherence score*. *Running time* menilai seberapa cepat model dapat memprediksi topik dari sebuah dokumen atau dataset. Evaluasi ini penting karena kecepatan prediksi menjadi pertimbangan utama saat model diterapkan dalam program.

Selain kecepatan, ketepatan prediksi model juga menjadi aspek penting. Salah satu cara untuk mengukur ketepatan atau kinerja sebuah model "*Topic Modelling*" adalah dengan mengukur topic coherence dari model tersebut (Kapadia, 2019). Konsep kohesi topik melibatkan penggabungan sejumlah matrik dalam sebuah kerangka kerja untuk mengevaluasi keterkaitan antar topik yang dihasilkan oleh model. Menurut Mimno (2011), *topic coherence* dapat dihitung dengan membandingkan pasangan kata-kata dalam suatu topik; semakin tinggi nilai kohesi, semakin baik modelnya (ArRosyid, 2022). *Coherence score* dapat dihitung menggunakan pendekatan *Intrinsic Umass Measure* sebagai berikut:

$$SCORE_{UMass}\,(W_i.W_j) = log\frac{D(W_i.W_j)+1}{D(W_i)} \qquad (1)$$

D = Total jumlah dokumen dalam korpus
D (Wi) = Total jumlah dokumen yang mengandung kata Wi
D (Wi.,Wj) = Total jumlah dokumen yang mengandung kata Wi dan Wj

HASIL DAN PEMBAHASAN

Data saran mahasiswa yang telah dikumpulkan disimpan dalam format CSV. Selanjutnya, data dalam format CSV tersebut akan diproses dalam tahap *preprocessing*. Karena data yang digunakan adalah teks tidak terstruktur, maka *preprocessing* yang dilakukan meliputi *case folding*, penghapusan angka, penghapusan emotikon, dan penghapusan kata-kata umum (*stopwords*). Hasil dari tahap *preprocessing* dapat dilihat pada tabel berikut.

Table 1 Hasil *Preprocessing* Data Saran Mahasiswa

Review	*Preprocessed review*
Saya berharap pelayanan akademik mampu memberikan pelayanan yang baik kepada mahasiswa dan seluruh masyarakat Unismuh.Terutama	berharap pelayanan akademik pelayanan mahasiswa

untuk para bapak/ibu dosen sekiranya mampu memberikan metode yang baik bagi mahasiswa sehingga mahasiswa lebih aktif dan produktif dalam proses pembelajaran.	masyarakat unismuh terutama dosen metode mahasiswa mahasiswa aktif produktif proses pembelajaran
Meningkatkan ruang lingkup belajar bagi mahasiswa seni ðŸ™‚ðŸ™‚»ðŸ™‚ðŸ»	Meningkatkan ruang lingkup belajar mahasiswa seni
SARAN SAYA YAITU AKAN LEBIH BAIK JIKA PERKULIAHAN DILAKUKAN SECARA LURING AGAR MAHASISWA LEBIH PAHAM DENGAN MATERI YANG DIBERIKAN	perkuliahan luring mahasiswa paham materi

Sebelumnya, data ulasan dipenuhi oleh berbagai gangguan seperti emoji, kata berulang, dan tanda baca yang berlebihan. Namun, setelah melalui tahap pembersihan, data kini bebas dari *noise* dan lebih terstruktur. Dengan data yang bersih dan terstruktur ini, analisis teks dapat dilakukan dengan lebih akurat dan efisien, menghasilkan temuan yang lebih bermakna.

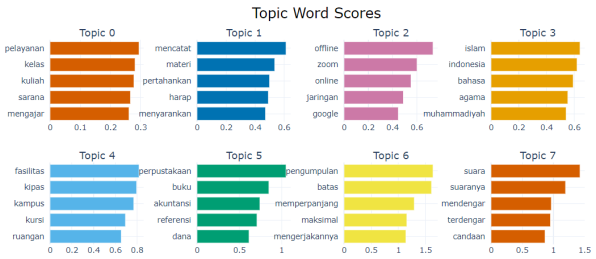
Proses BERTopic melibatkan beberapa komponen, seperti *SentenceTransformer* untuk menghasilkan *embedding*, UMAP untuk reduksi dimensi, HDBSCAN untuk *clustering*, dan CountVectorizer serta ClassTfidfTransformer untuk menghitung frekuensi kata dan transformasi TF-IDF berbasis kelas. Parameter yang digunakan mencakup model *embedding*, model UMAP, model HDBSCAN, model vectorizer, serta seed topics dan bahasa yang digunakan. Model menggunakan *SentenceTransformer* "*paraphrase-multilingual-MiniLM-L12-v2*" untuk menghasilkan *embedding*, kemudian menerapkan UMAP untuk reduksi dimensi menjadi 5 komponen. HDBSCAN digunakan untuk *clustering* dengan metrik euclidean dan metode pemilihan *cluster* "eom". CountVectorizer menghitung frekuensi kata, sementara ClassTfidfTransformer dengan pembobotan BM25 digunakan untuk transformasi TF-IDF berbasis kelas. Parameter lain termasuk daftar topik awal (*seed topics*), bahasa, jumlah topik, ukuran minimal topik, dan pilihan untuk menghitung probabilitas topik. Model ini menghasilkan topik dari teks dan menambahkan hasilnya ke dalam data.

Analisis dilakukan terhadap kata-kata dalam topik yang dihasilkan oleh algoritma BERTopic pada data saran mahasiswa di UNISMUH, menghasilkan 9 topik dari 26.009 saran. Topik-topik ini mencakup berbagai aspek kehidupan kampus, seperti pengajaran, fasilitas, administrasi, teknologi, kebersihan, dan keuangan. Algoritma menggunakan seed topics untuk memandu model dalam mengidentifikasi topik utama yang relevan, dengan setiap topik mencerminkan tema spesifik yang diwakili oleh kata-kata terkait.

	Topic	Count	Name
0	-1	9664	-1_kuliah_kelas_pelayanan_mata
1	0	11511	0_pelayanan_kelas_kuliah_sarana
2	1	1674	1_mencatat_materi_pertahankan_harap
3	2	1354	2_offline_zoom_online_jaringan
4	3	633	3_islam_indonesia_bahasa_agama
5	4	586	4_fasilitas_kipas_kampus_kursi
6	5	265	5_perpustakaan_buku_akuntansi_referensi
7	6	152	6_pengumpulan_batas_memperpanjang_maksimal
8	7	130	7_suara_suaranya_mendengar_terdengar
9	8	40	8_kegagalan_menyerah_memperoleh_berhenti

Gambar 3 Hasil Pengelompokan Berdasarkan Topik

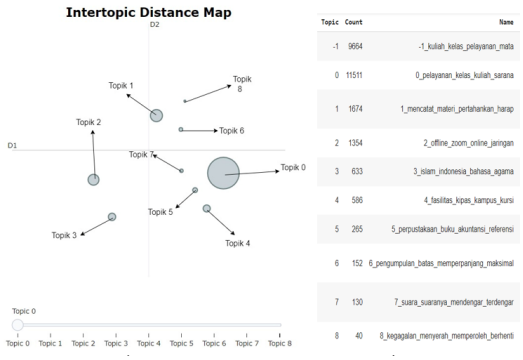
Setelah berhasil mengaplikasikan model topik, selanjutnya memvisualisasikan hasil pemodelan topik untuk mengevaluasi dan memahami topik-topik yang telah diekstrak.



Gambar 4 Pengelompikkan Berdasarkan Topik

Gambar tersebut menggambarkan distribusi kata-kata utama dalam 8 topik yang dihasilkan dari analisis data saran mahasiswa berdasarkan "*Topic Word Scores*." Setiap topik memiliki kata-kata yang sering muncul dan relevan, serta skor yang menunjukkan pentingnya kata tersebut dalam topik. Misalnya, Topik 0 terkait dengan layanan akademik dan sarana pembelajaran, dengan kata-kata seperti "pelayanan" dan "kelas." Topik 1 fokus pada aktivitas

mencatat dan saran, dengan kata-kata seperti "mencatat" dan "materi." Topik 2 mencakup aktivitas online dan jaringan, sementara Topik 3 berhubungan dengan Islam dan bahasa. Topik 4 terkait dengan fasilitas kampus, Topik 5 dengan ketersediaan referensi, Topik 6 dengan jadwal dan pengumpulan, dan Topik 7 berfokus pada suara. Visualisasi ini menampilkan topik-topik yang paling dominan, memudahkan interpretasi dengan hanya menyoroti topik yang paling relevan atau memiliki jumlah dokumen terbanyak.



Topic	Count	Name
-1	9664	-1_kuliah_kelas_pelayanan_mata
0	11511	0_pelayanan_kelas_kuliah_sarana
1	1674	1_mencatat_materi_pertahankan_harap
2	1354	2_offline_zoom_online_jaringan
3	633	3_islam_indonesia_bahasa_agama
4	586	4_fasilitas_kipas_kampus_kursi
5	265	5_perpustakaan_buku_akuntansi_referensi
6	152	6_pengumpulan_batas_memperpanjang_maksimal
7	130	7_suara_suaranya_mendengar_terdengar
8	40	8_kegagalan_menyerah_memperoleh_berhenti

Gambar 5 *Intertopic Distance Map* BERTopic

Visualisasi peta jarak antar topik dari korpus teks menunjukkan distribusi dan hubungan antar topik yang diwakili oleh lingkaran-lingkaran dengan berbagai ukuran dan posisi. Setiap lingkaran mewakili satu topik, dengan ukuran lingkaran menunjukkan frekuensi kemunculan topik tersebut dalam dataset. Topik 0 memiliki lingkaran terbesar dengan 11.511 dokumen, menandakan bahwa topik ini paling umum dan mencakup banyak aspek terkait layanan kuliah, seperti "pelayanan" dan "kelas." Jarak antara lingkaran menunjukkan kemiripan atau perbedaan konten antar topik. Misalnya, topik 4 dan 5 yang berhubungan dengan fasilitas kampus memiliki jarak dekat, menandakan kemiripan tema. Sebaliknya, topik 4 dan topik 8, yang berjauhan, menunjukkan perbedaan signifikan dalam tema, dengan topik 4 terkait fasilitas kampus dan topik 8 berfokus pada motivasi.

BERTopic membutuhkan waktu 13 menit untuk menyelesaikan pemodelan topik pada perangkat dengan prosesor Core i5-7300U, RAM 8GB, dan VRAM 4GB. Meskipun waktu ini cukup lama, hal tersebut dapat dimaklumi mengingat BERTopic menggunakan arsitektur *neural network* yang memerlukan pemrosesan intensif. Selain itu, BERTopic menghasilkan *coherence score* sebesar 0,6377.

```
coherence_score = calculate_coherence_score(model, df['saran_without_stopwords'])
print(f"Coherence Score: {coherence_score}")
```

```
Coherence Score: 0.637722900338999
```

Gambar 6 *Coherence Score* Algoritma BERTopic

Skor ini menunjukkan bahwa topik yang dihasilkan cukup koheren dan mudah diinterpretasikan. Dengan skor yang mendekati 1, BERTopic mampu mengidentifikasi topik yang tidak hanya relevan secara statistik tetapi juga semantik, sehingga menghasilkan tema yang logis dan mudah dipahami oleh manusia.

KESIMPULAN

Penerapan teknik pemodelan topik menggunakan BERTopic berhasil mengekstrak data saran mahasiswa dan menghasilkan sebanyak 9 topik dari total 26.009 saran. BERTopic membutuhkan waktu 13 menit dan menghasilkan *coherence score* sebesar 0.637. Hasil visualisasi menunjukkan bahwa topik-topik dalam kumpulan dokumen terdistribusi dalam beberapa kelompok utama, dengan topik dominan yang berkaitan dengan pengajaran dan layanan kelas, serta isu-isu terkait fasilitas kampus, sumber daya referensi, penggunaan teknologi, dan nilai-nilai agama. Berdasarkan temuan dari penelitian ini, disarankan agar studi berikutnya mengeksplorasi penerapan algoritma BERTopic pada berbagai jenis teks atau domain lain untuk mengevaluasi konsistensi kinerjanya di luar konteks data saran mahasiswa, karena pengujian dengan dataset yang berbeda dapat memberikan wawasan yang lebih mendalam mengenai keunggulan dan keterbatasan algoritma ini.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada rekan-rekan yang telah banyak membantu dalam menyelesaikan penyusunan penelitian ini, serta kepada semua pihak yang telah memberikan dukungan dan kontribusi berharga selama proses penelitian.

DAFTAR PUSTAKA

ArRosyid, A. T. S. (2022). Pemodelan Topik Terkait Pemerintah di Indonesia Menggunakan Metode Online Latent Dirichlet Allocation. In *Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta*.

Atagün, E., Hartoka, B., & Albayrak, A. (2021). Topic Modeling Using LDA and BERT Techniques: Teknofest Example. *Proceedings - 6th International*

Conference on Computer Science and Engineering, UBMK 2021, *September 2021*, 660–664. https://doi.org/10.1109/UBMK52708.2021.9558988

Kapadia, S. (2019). *Evaluate Topic Midels: Latent Dirichlet Allocation (LDA)*. Towards Data Science. https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

Kherwa, P., & Bansal, P. (2020). Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems*, *7*(24), 1–16. https://doi.org/10.4108/eai.13-7-2018.159623

Lopez-Martinez, R. E., & Sierra, G. (2020). Research trends in the international literature on natural language processing, 2000-2019 - A bibliometric study. *Journal of Scientometric Research*, *9*(3), 310–318. https://doi.org/10.5530/JSCIRES.9.3.38

Mustafac. (2021). *Bert For Topic Modeling (Bert vs LDA)*. Analytics Vidhya. https://medium.com/analytics-vidhya/bert-for-topic-modeling-bert-vs-lda-8076e72c602b