EKTRASI TOPIK ULASAN WARUNG KOPI TULUNGAGUNG MENGGUNAKAN INDOBERT

SKRIPSI

Oleh:
MUHAMMAD HAIKAL AZZADIN
NIM. 220605110157

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026

EKSTRASI TOPIK ULASAN WARUNG KOPI DI TULUNGAGUNG MENGGUNAKAN INDOBERT

SKRIPSI

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh :
MUHAMMAD HAIKAL AZZADIN
NIM. 220605110157

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2026

HALAMAN PERSETUJUAN

EKSTRASI TOPIK ULASAN WARUNG KOPI DI TULUNGAGUNG MENGGUNAKAN INDOBERT

SKRIPSI

Oleh :
MUHAMMAD HAIKAL AZZADIN
NIM. 220605110157

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 08 Juni 2026

Pembimbing I,	Pembimbing II,
Prof. Dr. Suhartono, M.Kom	H. Syahiduz Zaman, M.Kom
NIP. 19700502 200501 1 005	NIP. 19700502 200501 1 005

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Supriyono, M.Kom
NIP. 19841010 201903 1 012

HALAMAN PENGESAHAN

EKSTRASI TOPIK ULASAN WARUNG KOPI DI TULUNGAGUNG MENGGUNAKAN INDOBERT

SKRIPSI

Oleh :
MUHAMMAD HAIKAL AZZADIN
NIM. 220605110157

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 08 Juni 2026

Susunan Dewan Penguji

Ketua Penguji	: Dr. Zainal Abidin, M.Kom NIP. 19760613 200501 1 004	( )
Anggota Penguji I	: Nurizal Dwi Priandani, M.Kom NIP. 19920830 202203 1 001	( )
Anggota Penguji II	: Prof. Dr. Suhartono, M.Kom NIP. 19700502 200501 1 005	( )
Anggota Penguji III	: H. Syahiduz Zaman, M.Kom NIP. 19700502 200501 1 005	( )

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Supriyono, M.Kom
NIP. 19841010 201903 1 012

PERNYATAAN KEASLIAN TULISAN

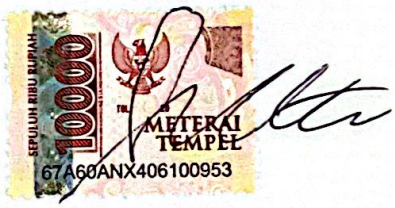
Saya yang bertanda tangan di bawah ini:

Nama	: MUHAMMAD HAIKAL AZZADIN
NIM	: 220605110157
Fakultas / Program Studi	: Sains dan Teknologi / Teknik Informatika
Judul Skripsi	: Ekstrasi topik ulasan warung kopi di Tulungagung Menggunakan indoBERT

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 18 Juni 2026
Yang membuat pernyataan,



MUHAMMAD HAIKAL AZZADIN
NIM. 220605110157

MOTTO

"Mereka menertawanku karena aku berbeda, aku menertawakan mereka karena mereka semua sama"

HALAMAN PERSEMBAHAN

Alhamdulillahi rabbil 'alamin, segala puji bagi Allah Swt. atas segala rahmat, nikmat, dan karunia-Nya yang senantiasa dilimpahkan kepada penulis. Shalawat serta salam semoga tetap tercurah kepada junjungan Nabi Muhammad saw., beserta keluarga, sahabat, dan seluruh umatnya hingga akhir zaman. Berkat rahmat, pertolongan, dan ridha Allah Swt., penulis diberikan kemudahan, kesehatan, serta kekuatan sehingga dapat menyelesaikan skripsi yang berjudul **"Ekstraksi Topik Ulasan Warung Kopi di Tulungagung Menggunakan BERTopic Berbasis IndoBERT"**.

Skripsi ini penulis persembahkan kepada kedua orang tua tercinta sebagai bentuk bakti, rasa hormat, cinta, dan terima kasih yang tiada terhingga atas segala kasih sayang, pengorbanan, doa, serta dukungan yang senantiasa diberikan. Doa dan perjuangan mereka menjadi sumber semangat dan kekuatan bagi penulis dalam menempuh pendidikan hingga mampu menyelesaikan skripsi ini.

Kepada saudara-saudara tercinta dan seluruh keluarga besar, penulis mempersembahkan karya sederhana ini sebagai ungkapan rasa syukur dan terima kasih atas perhatian, motivasi, dukungan, serta doa yang selalu menyertai setiap langkah perjalanan penulis dalam menyelesaikan pendidikan.

Kepada dosen pembimbing, para dosen, sahabat, teman-teman seperjuangan, serta semua pihak yang telah memberikan bantuan, arahan, dukungan, dan doa kepada penulis, penulis menyampaikan rasa terima kasih yang

sebesar-besarnya. Segala bentuk bantuan dan kebersamaan yang diberikan menjadi bagian penting dalam proses penyelesaian skripsi ini.

Semoga skripsi ini dapat memberikan manfaat, baik bagi penulis sendiri, dunia akademik, maupun masyarakat luas, serta menjadi amal baik yang bernilai ibadah di sisi Allah Swt. Aamiin ya Rabbal 'alamin.

KATA PENGANTAR

Alhamdulillāhirabbil ‘ālamīn, segala puji bagi Allah *Subḥānahu wa Ta‘ālā*. yang telah melimpahkan rahmat, taufik, hidayah, serta karunia-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul *“Ektrasi Topik ulasan warung kopi di Tulungagung Menggunakan indoBERT”* dengan baik. Shalawat serta salam semoga senantiasa tercurah kepada junjungan kita Nabi Muhammad *Shollallohu ‘Alaihi Wasallam*, beserta keluarga, sahabat, dan seluruh pengikutnya hingga akhir zaman. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer pada Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Prof. Dr. Hj. Ilfi Nurdiana, M.Si., CAHRM., CRMP., selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Dr. Agus Mulyono, M.Kes., selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Bapak Supriyono, M.Kom., selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Bapak Prof. Dr. Suhartono, M.Kom., selaku dosen pembimbing I dan Bapak H. Syahiduz Zaman, M.Kom., selaku dosen pembimbing II, yang telah memberikan bimbingan, arahan, masukan, motivasi, serta perhatian kepada penulis selama proses penyusunan skripsi ini. Terima kasih atas waktu,

kesabaran, dan dedikasi yang telah diberikan sehingga penelitian dan penulisan skripsi ini dapat terselesaikan dengan baik.

5. Bapak Dr. Zainal Abidin, M.Kom., selaku Ketua Penguji dan Ibu Nurizal Dwi Priandani, M.Kom., selaku Anggota Penguji I, yang telah memberikan evaluasi, kritik, saran, serta berbagai masukan yang membangun sehingga skripsi ini dapat disusun dengan lebih baik. Penulis mengucapkan terima kasih atas perhatian, arahan, dan ilmu yang diberikan selama proses ujian skripsi.
6. Seluruh dosen Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang, penulis menyampaikan ucapan terima kasih atas ilmu yang telah di berikan selama masa perkuliahan. Semoga segala ilmu dan kebaikan yang telah diberikan menjadi amal jariyah serta memberikan manfaat yang berkelanjutan bagi penulis di masa yang akan datang.
7. Ayah tercinta, Ali Mahmud, yang senantiasa menjadi teladan dalam kerja keras, keteguhan, dan tanggung jawab. Terima kasih atas doa, dukungan, nasihat, serta pengorbanan yang telah diberikan selama ini. Setiap perjuangan dan kasih sayang beliau menjadi sumber motivasi bagi penulis untuk terus berusaha hingga mampu menyelesaikan skripsi ini dengan baik.
8. Mama tercinta, Sujiyuana, yang senantiasa memberikan kasih sayang, doa, dukungan, motivasi, serta pengorbanan tanpa henti kepada penulis. Ketulusan, keikhlasan, dan doa yang tidak pernah putus menjadi kekuatan

bagi penulis dalam menghadapi setiap proses dan tantangan hingga skripsi ini dapat terselesaikan dengan baik.

9. Kakak tercinta, Arrizna Aufani dan Ikfina Akmalia, yang selalu memberikan doa, semangat, perhatian, serta dukungan kepada penulis dalam setiap proses perjalanan pendidikan hingga penyusunan skripsi ini.
10. Seluruh keluarga besar penulis yang tidak dapat disebutkan satu per satu, terima kasih atas doa, dukungan, perhatian, dan semangat yang telah diberikan kepada penulis selama menempuh pendidikan.
11. Teman-teman kontrakan, yaitu Eco, Ijun, Mahdy, Asror, dan Rizky, yang telah menjadi keluarga kedua selama masa perkuliahan. Terima kasih atas kebersamaan, canda tawa, bantuan, serta dukungan yang menjadikan perjalanan akademik ini lebih berkesan.
12. Sahabat seperjuangan, yaitu Andre, Annas, Ageng, Padli, Ega, Najib, Hasyim, Nizam, Akbar, Jidan, dan Rizky yang tergabung dalam Grup Takmir Gengs, yang senantiasa memberikan semangat, motivasi, dukungan, serta kebersamaan kepada penulis selama menjalani masa perkuliahan hingga proses penyusunan skripsi ini. Terima kasih atas setiap canda tawa, kebersamaan, bantuan, dan pengalaman berharga yang telah menjadi bagian dari perjalanan akademik penulis.
13. Teman-teman di rumah yang tergabung dalam Grup Aduh Terenyuh, yang selalu memberikan perhatian, doa, semangat, dan dukungan kepada penulis, baik dalam suka maupun duka. Terima kasih karena selalu hadir menjadi tempat berbagi cerita dan saling menguatkan.

14. Ucapan terima kasih penulis sampaikan kepada keluarga besar Teknik Informatika angkatan 2022 Infinity, yang telah memberikan kebersamaan, dukungan, dan pengalaman berharga selama masa perkuliahan.
15. Terakhir, kepada diri penulis sendiri, terima kasih telah bertahan, berjuang, dan tidak menyerah meskipun harus melewati berbagai tantangan, keraguan, dan kelelahan selama perjalanan perkuliahan hingga penyusunan skripsi ini. Terima kasih karena telah percaya bahwa setiap proses akan menemukan akhirnya. Semoga segala usaha yang telah dilakukan menjadi bekal untuk terus melangkah, belajar, dan memberikan manfaat bagi banyak orang.

Penulis menyadari bahwa skripsi ini masih jauh dari kata sempurna. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan demi perbaikan di masa mendatang. Semoga skripsi ini dapat memberikan manfaat bagi pembaca dan menjadi referensi bagi penelitian selanjutnya.

Malang, 13 Juni 2026

Penulis

DAFTAR ISI

HALAMAN PENGAJUAN ..... **ii**
HALAMAN PERSETUJUAN ..... **iii**
HALAMAN PENGESAHAN ..... **iv**
PERNYATAAN KEASLIAN TULISAN ..... **v**
MOTTO ..... **vi**
HALAMAN PERSEMBAHAN ..... **vii**
KATA PENGANTAR ..... **ix**
DAFTAR ISI ..... **xiii**
DAFTAR GAMBAR ..... **xv**
DAFTAR TABEL ..... **xvi**
ABSTRAK ..... **xvii**
ABSTRACT ..... **xviii**
البحث مستخلص ..... **xix**
BAB I PENDAHULUAN **1**
1.1 Latar Belakang 1
1.2 Pernyataan Masalah 5
1.3 Tujuan Penelitian 6
1.4 Batasan Masalah 6
1.5 Manfaat Penelitian 7
BAB II STUDI PUSTAKA **9**
2.1 Penelitian Terkait 9
2.2 Pemrosesan Bahasa Alami (*Natural Languange processing*) 16
2.3 Analisis Topik Pada Ulasan pengunjung 18
2.4 Analisis Topik Pada Data Teks 21
2.5 Model bahasa Pra-latih indoBERT 23
2.6 BERTopic 25
BAB III DESAIN DAN IMPLEMENTASI SISTEM **28**
3.1 Desain Penelitian 28
3.1.1 Pengumpulan Data 29
3.1.2 Desain Sistem 31
3.1.3 Preprocessing Data 31
3.1.4 Representasi Teks indoBERT 36
3.1.5 Proses Analisis Topik 37
3.1.6 Topic coherence 46
3.2 Implementasi Penelitian 48
3.3 Implementasi Representasi Teks Dengan IndoBERT 54
3.4 Implementasi Proses Analisis Topik 55
BAB IV HASIL DAN PEMBAHASAN **64**
4.1 Data Penelitian 64

4.2 Hasil Pemodelan topik BERTopic 64
4.3 Visualisasi Topik 66
4.3.1 Intertopic Distance Map 67
4.3.2 Topic Word Scores 68
4.3.3 WordCloud 71
4.3.4 Distribusi Topik 77
4.3.5 Topic Coherence 79
4.3.6 Perbandingan BERTopic dan LDA 81
4.4 Pembahasan 85
BAB V KESIMPULAN DAN SARAN 89
5.1 Kesimpulan 89
5.2 Saran 90
DAFTAR PUSTAKA

DAFTAR GAMBAR

Gambar 3. 1 prosedur penelitian 28
Gambar 3. 2 Desain Sistem 32
Gambar 3. 3 Proses indoBERT 37
Gambar 3. 4 Intertopic Distance 40
Gambar 3. 5 Reduksi UMAP 42
Gambar 3. 6 HDBSCAN 43
Gambar 3. 7 Topic Coherence 47
Gambar 3. 8 Sourcode Tahap Cleaning 48
Gambar 3. 9 Sourecode Case Folding 49
Gambar 3. 10 Source code Normalisasi kata 50
Gambar 3. 11Sourcode Stopword Removal 51
Gambar 3. 12 Source code Tokenizing 52
Gambar 3. 13 Source code BERTopic 54
Gambar 3. 14 Alur BERTopic 55
Gambar 3. 15 Source ocode BERTopic 56
Gambar 4. 1 Hasil Inttertopic Distance Map 67
Gambar 4. 2 Hasil Topic Word Scores 69
Gambar 4. 3 Topic 0 71
Gambar 4. 4 Topic 1 72
Gambar 4. 5 Hasil Topik 2 73
Gambar 4. 6 Hasil Topik 3 74
Gambar 4. 7 Hasil Topik 4 75
Gambar 4. 8 Hasil Topik 5 76
Gambar 4. 9 Distribusi Topik 77
Gambar 4. 10 Topic Coherence BERTopic 80
Gambar 4. 11 Perbandingan Score Coherence 82
Gambar 4. 12 Perbandingan Score Coherence Per Topik 83

DAFTAR TABEL

Tabel 2. 1 Penelitian Terdahulu 9
Tabel 3. 1 Pengumpulan Data 30
Tabel 3. 2 Proses Cleaning 33
Tabel 3. 3 Proses Case Folding 34
Tabel 3. 4 Proses Normalisasi Kata 34
Tabel 3. 5 proses Stopword Removal 35
Tabel 3. 6 Proses Tokenizing 36
Tabel 3. 7 Final Preprocessing 53
Tabel 3. 8 Contoh data topik 59
Tabel 3. 9 Frekuensi kata 59
Tabel 3. 10 Perhitungan TF 60
Tabel 3. 11 Nilai Global 60
Tabel 3. 12 Perhitungan ICF 61
Tabel 3. 13 Perhitungan c-TF-IDF 61
Table 4. 1 Data penelitian 64
Table 4. 2 Informasi Topik 65
Table 4. 3 Nilai Coherence BERTopic 80
Table 4. 4 Perbandingan Coherence Global Model BERTopic dan LDA 81
Table 4. 5 Perbandingan Coherence pertopik 84

ABSTRAK

Azzadin, Muhammad Haikal Azzadin. 2026. **Ekstrasi Topik Ulasan Warung kopi di Tulungagung Menggunakan INDOBERT.** Skripsi. Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Prof. Dr. Suhartono, M.Kom (II) H. Syahiduz Zaman, M.Kom

.

Kata Kunci: Ekstrasi Topik, BERTopic, LDA, Topic Coherence

Ulasan pelanggan pada platform Google Maps menjadi sumber informasi yang penting untuk mengetahui pengalaman dan preferensi masyarakat terhadap warung kopi di Tulungagung. Banyaknya ulasan yang tersedia dalam bentuk teks menyebabkan proses identifikasi topik secara manual menjadi kurang efektif sehingga diperlukan metode otomatis untuk melakukan ekstraksi topik. Penelitian ini bertujuan menerapkan metode BERTopic berbasis IndoBERT dalam melakukan ekstraksi topik ulasan warung kopi di Tulungagung serta membandingkan kualitas topik yang dihasilkan dengan metode Latent Dirichlet Allocation (LDA). Data penelitian diperoleh dari ulasan Google Maps warung kopi di Tulungagung yang melalui tahapan preprocessing, representasi teks menggunakan IndoBERT, pemodelan topik menggunakan BERTopic, serta evaluasi menggunakan topic coherence (C_v). Sebagai pembanding, dilakukan pemodelan topik menggunakan metode LDA pada data yang sama. Hasil penelitian menunjukkan bahwa BERTopic berbasis IndoBERT mampu menghasilkan enam topik utama yang merepresentasikan tema-tema dalam ulasan pelanggan. Berdasarkan evaluasi menggunakan topic coherence, BERTopic memperoleh nilai coherence global sebesar 0,5859, sedangkan LDA memperoleh nilai sebesar 0,4226 serta menghasilkan nilai coherence yang lebih tinggi pada lima dari enam topik yang terbentuk. Temuan ini menunjukkan bahwa BERTopic berbasis IndoBERT mampu menghasilkan topik yang lebih koheren dan mudah diinterpretasikan dibandingkan LDA, sehingga berpotensi menjadi metode yang efektif untuk ekstraksi topik pada data ulasan berbahasa Indonesia.

ABSTRACT

Azzadin, Muhammad Haikal Azzadin. 2026. **Ekstrasi Topik Ulasan Warung kopi di Tulungagung Menggunakan INDOBERT.** Skripsi. Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Prof. Dr. Suhartono, M.Kom (II) H. Syahiduz Zaman, M.Kom

Key words: Topic Extraction, BERTopic, LDA, Topic Coherence

Customer reviews on Google Maps serve as an important source of information for understanding people's experiences and preferences regarding coffee shops in Tulungagung. The large volume of text-based reviews makes manual topic identification inefficient, creating the need for an automated topic extraction method. This study aims to implement the BERTopic method based on IndoBERT for topic extraction from coffee shop reviews in Tulungagung and to compare its performance with the Latent Dirichlet Allocation (LDA) method. The dataset was collected from Google Maps reviews of coffee shops in Tulungagung and processed through several stages, including preprocessing, text representation using IndoBERT embeddings, topic modeling with BERTopic, and evaluation using topic coherence (C_v). As a baseline model, LDA was applied to the same dataset for comparison. The results show that BERTopic based on IndoBERT successfully generated six main topics representing the themes contained in customer reviews. Based on topic coherence evaluation, BERTopic achieved a global coherence score of 0.5859, while LDA obtained a score of 0.4226. In addition, BERTopic produced higher coherence scores in five out of the six identified topics. These findings indicate that BERTopic based on IndoBERT generates more coherent and interpretable topics than LDA, making it a promising approach for topic extraction from Indonesian text reviews.

البحث مستخلص

عزادين، محمد هيكل عزادين. 2026. **استخراج الموضوعات من مراجعات المقاهي في تولونغاغونغ باستخدام نموذج إندوبيرت** (IndoBERT). رسالة جامعية. قسم تقنية المعلومات، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية مالانغ. المشرف الأول: الأستاذ الدكتور سهارتونو، الماجستير. المشرف الثاني: الحاج شَهِيد الزمان، الماجستير.

الكلمات المفتاحية :استخراج الموضوعات، BERTopic، LDA، اتساق الموضوعات (Topic Coherence)..

تُعد مراجعات العملاء على منصة Google Maps مصدرًا مهمًا للمعلومات لفهم تجارب المجتمع وتفضيلاته تجاه المقاهي في تولونغاغونغ. ونظرًا لكثرة المراجعات النصية، فإن تحديد الموضوعات يدويًا يصبح أقل كفاءة، مما يستدعي استخدام أساليب آلية لاستخراج الموضوعات. تهدف هذه الدراسة إلى تطبيق طريقة BERTopic المعتمدة على IndoBERT لاستخراج الموضوعات من مراجعات المقاهي في تولونغاغونغ، ومقارنة جودة الموضوعات الناتجة مع طريقة التخصيص الكامن لديريشليه.(LDA) تم جمع بيانات الدراسة من مراجعات Google Maps الخاصة بالمقاهي في تولونغاغونغ، ثم خضعت لمراحل المعالجة المسبقة للنصوص، وتمثيل النصوص باستخدامIndoBERT ، ونمذجة الموضوعات باستخدام BERTopic، وتقييم النتائج باستخدام مقياس اتساق الموضوعات .(C_v) كما استُخدمت طريقة LDA كنموذج مقارن على البيانات نفسها.أظهرت نتائج الدراسة أن BERTopic المعتمد على IndoBERT تمكن من استخراج ستة موضوعات رئيسية تمثل الموضوعات المختلفة الواردة في مراجعات العملاء. وبناءً على تقييم اتساق الموضوعات، حقق BERTopic قيمة اتساق كلية بلغت 0.5859، في حين حقق LDA قيمة 0.4226، كما تفوق BERTopic في خمسة من أصل ستة موضوعات مستخرجة. وتشير هذه النتائج إلى أن BERTopicالمعتمد على IndoBERT قادر على إنتاج موضوعات أكثر اتساقًا وأسهل في التفسير مقارنةً بـLDA ، مما يجعله أسلوبًا واعدًا وفعالًا لاستخراج الموضوعات من النصوص الإندونيسية.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Revolusi digital yang terjadi dalam zaman sekarang telah mengubah pandagan kehidupan masyarakat secara fundamental, termasuk dalam cara mereka mengonsumsi informasi dan membuat keputusan pembelian. Di ranah ekonomi kreatif, subsektor kuliner mengalami transformasi yang cukup masif dengan hadirnya platform digital berbasis geolokasi dan ulasan pengguna, seperti Google Maps. Platform semacam ini memungkinkan setiap pengguna untuk memberikan rating sekaligus ulasan tertulis mengenai pengalaman mereka di suatu tempat usaha. Data ulasan ini, jika digali lebih dalam, menyimpan informasi berharga—mulai dari kepuasan pelanggan, keluhan, hingga tren preferensi masyarakat. Calon pelanggan pun sering kali menjadikan ulasan tersebut sebagai bahan pertimbangan utama sebelum mengunjungi suatu tempat. Salah satu jenis usaha yang sangat dipengaruhi oleh fenomena ini adalah warung kopi, yang saat ini menjadi primadona bisnis kuliner di berbagai wilayah Indonesia, termasuk di Kabupaten Tulungagung. Warkop tidak hanya berfungsi sebagai tempat menikmati minuman kopi, tetapi secara pandangan lain bisa digunakan untuk ruang sosial untuk berkumpul, berdiskusi, dan beraktivitas. Seiring dengan meningkatnya jumlah warung kopi, persaingan antarpelaku usaha juga semakin ketat. Oleh karena itu, pemahaman terhadap persepsi dan pengalaman pengunjung menjadi faktor penting dalam meningkatkan kualitas layanan dan mempertahankan daya saing usaha (Filieri & others, 2021).

Dalam konteks industri kuliner, khususnya warung kopi, keberadaan ulasan pelanggan menjadi indikator penting dalam menilai kualitas layanan, produk, maupun suasana tempat. Warung kopi tidak lagi sekadar tempat konsumsi, tetapi juga menjadi ruang sosial yang mendukung interaksi masyarakat. Seiring meningkatnya jumlah usaha kuliner, persaingan antar pelaku usaha menjadi semakin ketat, sehingga pemahaman terhadap persepsi pelanggan menjadi kebutuhan yang krusial. Namun, banyaknya jumlah ulasan yang tersedia seringkali menyulitkan pelaku usaha untuk menganalisis informasi secara manual, terutama ketika jumlah data mencapai ratusan hingga ribuan ulasan.

Dalam perspektif Islam, pentingnya memahami dan memverifikasi informasi sebelum mengambil keputusan juga telah dijelaskan dalam Al-Qur'an. Allah SWT berfirman dalam Surah Al-Hujurat ayat 6:

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْٓا اِنْ جَاۤءَكُمْ فَاسِقٌۢ بِنَبَاٍ فَتَبَيَّنُوْٓا اَنْ تُصِيْبُوْا قَوْمًاۢ بِجَهَالَةٍ فَتُصْبِحُوْا عَلٰى مَا فَعَلْتُمْ نٰدِمِيْنَ

"*Wahai orang-orang yang beriman, jika seorang fasik datang kepadamu membawa berita penting, maka telitilah kebenarannya agar kamu tidak mencelakakan suatu kaum karena ketidaktahuan(-mu) yang berakibat kamu menyesali perbuatanmu itu."(Q.S Al Hujarat: 6).*

Selain itu, Islam juga menganjurkan umatnya untuk menjaga dan mempererat hubungan sosial melalui silaturahmi. Rasulullah SAW bersabda bahwa "*siapa yang ingin dilapangkan rezekinya dan dipanjangkan umurnya maka hendaklah ia menyambung silaturahmi*" (*HR. Bukhari No. 5986 dan Muslim No. 2557).*

Ayat dan hadis tersebut menunjukkan bahwa dalam kehidupan sosial, manusia tidak hanya dituntut untuk menjalin hubungan yang baik melalui interaksi dan komunikasi, tetapi juga harus mampu memilah serta memahami informasi yang diterima secara bijak. Dalam konteks kehidupan modern, aktivitas interaksi sosial tersebut dapat terjadi di berbagai tempat, salah satunya adalah warung kopi yang kini tidak hanya berfungsi sebagai tempat menikmati minuman, tetapi juga sebagai ruang komunikasi, diskusi, dan pertukaran informasi antar individu.

Seiring dengan berkembangnya teknologi digital, ulasan pengunjung pada platform seperti Google Maps menjadi salah satu sumber informasi yang penting dalam menilai kualitas suatu tempat. Namun, banyaknya data ulasan yang tersedia seringkali menyulitkan dalam proses analisis secara manual. Oleh karena itu, diperlukan suatu pendekatan yang mampu mengolah dan mengekstraksi informasi secara sistematis, akurat, dan objektif, Sebagian besar penelitian terdahulu menggunakan data Twitter, Instagram, dan X yang memiliki karakteristik teks pendek dan dinamis. Sementara itu, ulasan Google Maps memiliki karakteristik berbeda karena berisi pengalaman langsung pelanggan terhadap layanan dan produk. Karena hal ini diperlukan penelitian yang khusus membahas mengenai eksplorasi topik ulasan Google Maps menggunakan pendekatan BERTopic berbasis IndoBERT.

Untuk mengatasi permasalahan tersebut, pendekatan berbasis Natural Language Processing (NLP) dapat digunakan untuk mengekstraksi informasi dari data teks secara otomatis. Salah satu model yang banyak digunakan dalam pemrosesan bahasa Indonesia adalah IndoBERT, yaitu model berbasis transformer

yang mampu memahami konteks semantik teks dengan lebih baik dibandingkan metode tradisional. Penelitian sebelumnya menunjukkan bahwa IndoBERT memiliki performa yang unggul dalam berbagai tugas NLP seperti analisis sentimen dan klasifikasi teks berbahasa Indonesia (Koto et al., 2020).

Meskipun demikian, sebagian besar penelitian yang memanfaatkan IndoBERT masih berfokus pada pendekatan supervised learning seperti klasifikasi sentimen yang memerlukan data berlabel. Pendekatan ini memiliki keterbatasan dalam menggali informasi yang lebih luas, seperti topik pembahasan utama yang muncul dalam ulasan. Meskipun demikian, sebagian besar penelitian yang memanfaatkan IndoBERT masih berfokus pada pendekatan supervised learning seperti klasifikasi sentimen yang memerlukan data berlabel. Pendekatan ini memiliki keterbatasan dalam menggali informasi yang lebih luas, seperti topik pembahasan utama yang muncul dalam ulasan. Oleh karena itu, diperlukan pendekatan unsupervised learning yang mampu mengelompokkan data secara otomatis tanpa membutuhkan label. Salah satu alternatif yang dapat dimanfaatkan adalah BERTopic, yaitu metode topic modeling berbasis embedding yang mengintegrasikan representasi semantik dari arsitektur BERT dengan algoritma clustering dan c-TF-IDF untuk membentuk topik-topik yang lebih mudah diinterpretasikan (Grootendorst, 2022). BERTopic terbukti mampu menghasilkan topik yang dapat membantu dalam mengidentifikasi informasi yang relevan.

Beberapa penelitian telah mengombinasikan model berbasis BERT dengan metode topic modeling untuk menganalisis opini publik seperti yang dilakukan pada penelitian yang dibahas oleh (Egger & others, 2022) yang Menunjukkan

bahwa penggunaan embedding berbasis transformer dalam topic modeling mampu meningkatkan kualitas interpretasi topik pada data ulasan online. Selain itu, penelitian terbaru di Indonesia juga mulai menggabungkan IndoBERT dengan BERTopic untuk menganalisis opini publik di media sosial dan platform ulasan, yang menunjukkan bahwa pendekatan ini efektif dalam mengidentifikasi pola diskusi dan persepsi masyarakat secara lebih mendalam. namun demikian, berdasarkan kajian terhadap penelitian terdahulu, masih terdapat beberapa keterbatasan. Sebagian besar penelitian masih berfokus pada data media sosial seperti Twitter dan Instagram, sementara pemanfaatan data ulasan Google Maps, khususnya pada sektor kuliner lokal, masih relatif terbatas. Selain itu, penelitian yang mengombinasikan IndoBERT dengan BERTopic dalam konteks analisis ulasan berbahasa Indonesia juga masih belum banyak dilakukan. Di sisi lain, evaluasi pada pendekatan unsupervised seperti topic modeling seringkali belum dibahas secara mendalam, khususnya dalam penggunaan metrik topic coherence sebagai indikator kualitas topik.

1.2 Pernyataan Masalah

Dengan mempertimbangkan latar belakang yang telah diuraikan, permasalahan yang diajukan dalam penelitian ini adalah sebagai berikut:

1. Bagaimana perenapan model indoBERT sebagai representasi teks ulasan pengunjung warung kopi waris di Tulungagung?
2. Bagaimana hasil pemodelan topik menggunakan metode BERTopic berbasis indoBERT pada ulasan pengunjung warung kopi waris di Tulungagung?

3. Bagaimana performa metode BERTopic dibandingankan dengan metode LDA berdasarkan metik Topic coherence ?

1.3 Tujuan Penelitian

Penelitian ini menetapkan beberapa batasan agar fokus analisis lebih terarah. Adapun batasan masalah yang digunakan dalam penelitian ini adalah sebagai berikut:

1. Penelitian hanya berfokus pada analisis topik tanpa melakukan analisis sentimen atau klasifikasi teks.
2. Representasi teks ulasan dilakukan menggunakan model bahasa IndoBERT (indobert-base-p2) dalam kondisi pretrained untuk menghasilkan embedding semantik, tanpa modifikasi arsitektur maupun proses fine-tuning.
3. Data yang digunakan berupa ulasan pengunjung warung kopi waris di Tulungagung yang diperoleh dari Google Maps.
4. Metode pemodelan topik yang digunakan adalah BERTopic, dengan Latent Dirichlet Allocation (LDA) sebagai metode pembanding.
5. Evaluasi kualitas topik dilakukan menggunakan metrik topic coherence (Cv).

1.4 Batasan Masalah

Adapun tujuan dari penelitian ini mencakup hal-hal berikut:

1. Menerapkan model IndoBERT untuk merepresentasikan teks ulasan pengunjung warung kopi Waris di Tulungagung ke dalam bentuk embedding numerik yang mampu menangkap makna semantik.
2. Mengimplementasikan metode BERTopic serta membandingkannya dengan metode LDA dalam mengidentifikasi dan mengelompokkan topik-topik utama yang muncul dalam ulasan pengunjung warung kopi di Tulungagung.
3. Mengevaluasi bagaimana kualitas topik yang dihasilkan oleh model menggunakan metrik topic coherence.

1.5 Manfaat Penelitian

Manfaat penelitian diharapkan dapat mencapai beberapa aspek berikut:

1. Memberikan kontribusi sebagai referensi dalam bidang Machine Learning dan Natural Language Processing (NLP), khususnya pada analisis topik berbasis unsupervised learning dengan memanfaatkan IndoBERT dan BERTopic.
2. Memberikan gambaran mengenai pemanfaatan embedding teks berbasis transformer dalam mengidentifikasi pola topik pada data teks tidak terstruktur, khususnya pada ulasan Google Maps.
3. Menjadi acuan bagi peneliti selanjutnya dalam pengembangan penelitian terkait analisis topik, text mining, serta pemanfaatan IndoBERT pada data berbahasa Indonesia.
4. Penelitian ini diharapkan dapat membantu pelaku usaha, khususnya pemilik warung kopi di Tulungagung, dalam memahami topik-topik utama yang

sering dibahas oleh pengunjung melalui ulasan, sehingga dapat digunakan sebagai bahan evaluasi dan peningkatan kualitas layanan.

BAB II

STUDI PUSTAKA

2.1 Penelitian Terkait

Untuk menyusun kerangka teoritis yang solid dan merancang metodologi yang sesuai, dilakukan penelusuran literatur terhadap berbagai studi yang memiliki korelasi dengan penelitian ini. Secara spesifik, kajian difokuskan pada penelitian-penelitian di bidang Natural Language Processing (NLP) yang mengimplementasikan arsitektur transformer—terutama IndoBERT—serta pendekatan topic modeling berbasis klaster seperti BERTopic. Berdasarkan hasil telaah, ditemukan bahwa penelitian sebelumnya umumnya terbagi menjadi dua aliran utama: pertama, pendekatan supervised learning untuk tugas klasifikasi teks; kedua, pendekatan unsupervised learning untuk eksplorasi topik-topik laten dalam korpus dokumen. Pemetaan terhadap beragam pendekatan ini menjadi pijakan awal bagi peneliti untuk menentukan metode yang paling sesuai dengan karakteristik data dan tujuan penelitian yang ingin dicapai. Rangkuman dari studi-studi tersebut disajikan pada Tabel 2.1 berikut ini.

Tabel 2. 1 Penelitian Terdahulu

No	Judul	Hasil	Relevansi
1.	Utilizing IndoBERT and BERTopic to Explore Public Opinion on BPS Instagram Posts (IndoBERT + BERTopic)	Menghasilkan topik opini publik yang kontekstual, namun terbatas pada media sosial. Hasil analisis sentimen menunjukkan bahwa dari 3.075. komentar yang	Penelitian ini sangat relevan karena menggunakan kombinasi IndoBERT dan BERTopic dalam analisis opini publik, sama seperti penelitian yang dilakukan. Hal ini mendukung penggunaan embedding IndoBERT dan metode BERTopic dalam menghasilkan

No	Judul	Hasil	Relevansi
		dianalisis, komentar netral mendominasi (52,78%), diikuti oleh komentar negatif (31,54%), dan positif (15,69%). Dengan rata score coherence 0,4152	topik yang kontekstual pada data teks.
2.	Sentiment Analysis and Topic Modelling Using IndoBERTweet and BERTopic for Public Health Issues (IndoBERTweet + BERTopic)	kombinasi IndoBERTweet dan BERTopic mampu menghasilkan kinerja yang baik dengan nilai Weighted F1-Score sebesar 0.7822, dengan nilai *coherance score* 7,154 Pemodelan topik menghasilkan 44 topik dengan 9 topik utama	Penelitian ini relevan karena menggunakan pendekatan BERTopic berbasis embedding transformer untuk pemodelan topik. Hal ini mendukung penggunaan metode BERTopic dalam penelitian ini, meskipun menggunakan varian model yang berbeda yaitu IndoBERT
3.	Analysis of public opinion on the governor candidate debate using LDA and IndoBERT (IndoBERT + LDA)	Penelitian ini menganalisis opini publik terhadap debat calon gubernur menggunakan metode LDA untuk pemodelan topik dan IndoBERT untuk klasifikasi sentimen. Hasil evaluasi menunjukkan nilai perplexity sebesar -7.7909 dan coherence score sebesar 0.5109 dengan 5 topik yang berhasil diidentifikasi.	Penelitian ini relevan karena sama-sama menggunakan IndoBERT dalam analisis teks. Perbedaannya terletak pada metode pemodelan topik yang digunakan, dimana penelitian ini menggunakan LDA, sedangkan penelitian yang dilakukan menggunakan BERTopic yang lebih modern dan berbasis embedding.
4.	Sentiment Analysis of Cyber Attacks in Bank Syariah	Pengujian model menunjukkan	Penelitian ini relevan karena menggunakan IndoBERT sebagai

No	Judul	Hasil	Relevansi
	Indonesia Using SVM and Indobert Methode (IndoBERT)	bahwa IndoBERT memiliki performa yang lebih baik dibandingkan SVM dalam mengklasifikasikan sentimen. Dengan akurasi 85% dan F1-Score 82%, IndoBERT terbukti lebih efektif dalam menangani berbagai jenis sentimen.	model representasi teks. Hal ini mendukung penggunaan IndoBERT dalam penelitian ini, meskipun fokus penelitian berbeda yaitu pada analisis sentimen.
5.	An optimized BERTopic framework based on cluster silhoue?e for improving topic coherence (BERTopic)	Mengusulkan optimasi pada metode BERTopic dengan memanfaatkan hubungan antara silhouette score dan topic coherence (c_NPMI), serta mengembangkan mekanisme optimasi dinamis untuk meningkatkan kualitas topik. Hasil eksperimen pada tiga dataset menunjukkan adanya peningkatan rata-rata coherence sebesar 12,24%, yang menandakan peningkatan kemampuan BERTopic dalam pembelajaran topik secara unsupervised.	Penelitian ini sangat relevan karena membahas pengembangan dan optimasi metode BERTopic dalam meningkatkan kualitas topik. Hal ini menjadi dasar dalam penelitian ini dalam memahami performa dan evaluasi kualitas topik yang dihasilkan.
6.	Analisis Perbandingan Model CNN dan IndoBERT Dalam Sentimen Berita Politik Indonesia (CNN + indoBERT)	membandingkan model CNN dan IndoBERT dalam klasifikasi	Penelitian ini relevan karena menunjukkan keunggulan IndoBERT dalam memahami teks berbahasa Indonesia. Hal ini

No	Judul	Hasil	Relevansi
		sentimen berita politik berbahasa Indonesia. Hasil evaluasi menunjukkan bahwa IndoBERT memiliki performa yang lebih unggul, dengan akurasi validasi tertinggi sebesar 92.93% dan metrik evaluasi yang stabil di semua kelas. Sementara itu, CNN mencatat akurasi 89.13%,	mendukung penggunaan IndoBERT dalam penelitian ini sebagai representasi teks sebelum dilakukan pemodelan topik.
7.	Analisis Perbandingan Klasifikasi Intent Chatbot Menggunakan Deep Learning BERT, RoBERTa, dan IndoBERT (BERT, ROBERTA, INDOBERT)	klasifikasi intent chatbot menggunakan model BERT, RoBERTa dan IndoBERT menghasilkan kesimpulan bahwa model BERT, RoBERTa dan IndoBERT dapat digunakan secara efektif untuk klasifikasi intent chatbot berbahasa Indonesia. Setelah menerjemahkan dataset chatbot universitas ke dalam Bahasa Indonesia, model IndoBERT mencapai akurasi 0.94, mengungguli model BERT yang mencapai 0.89 dan model RoBERTa yang mencapai 0.84. Hal ini menunjukkan bahwa performa IndoBERT lebih	Penelitian ini relevan karena membandingkan performa model berbasis transformer termasuk IndoBERT dalam memahami teks bahasa Indonesia. Hal ini memperkuat pemilihan IndoBERT dalam penelitian ini sebagai model embedding yang efektif.

No	Judul	Hasil	Relevansi
		baik dalam klasifikasi intent chatbot dibandingkan BERT dan RoBERTa	

Penelitian oleh (Anugrah & Agatha, 2025) mengusulkan pemanfaatan IndoBERT dan BERTopic untuk mengeksplorasi opini publik pada komentar Instagram. IndoBERT digunakan untuk menghasilkan embedding semantik yang kemudian diproses menggunakan BERTopic dalam mengidentifikasi topik-topik utama. Hasil penelitian menunjukkan bahwa metode tersebut mampu menghasilkan beberapa topik yang merepresentasikan persepsi masyarakat secara kontekstual dan mudah diinterpretasikan. Namun, penelitian ini masih terbatas pada data media sosial.

Selanjutnya, penelitian oleh (Sihombing & Widiyaningtyas, 2025) menggunakan IndoBERTweet dan BERTopic untuk menganalisis opini publik pada media sosial X. Pendekatan ini memanfaatkan embedding berbasis transformer untuk meningkatkan kualitas representasi teks sebelum dilakukan pemodelan topik. Hasil penelitian menunjukkan bahwa model mampu mengidentifikasi sejumlah topik utama yang mencerminkan isu dominan yang dibahas oleh pengguna. Namun, penelitian ini belum mengkaji data ulasan berbasis lokasi seperti Google Maps.

Penelitian oleh (Chamid et al., 2025) Mengombinasikan penggunaan IndoBERT dengan metode Latent Dirichlet Allocation dalam analisis topik pada

data teks. Hasil penelitian menunjukkan bahwa penggunaan IndoBERT mampu meningkatkan kualitas representasi teks sebelum dilakukan pemodelan topik. Namun, metode LDA masih memiliki keterbatasan dalam menangkap konteks semantik secara mendalam dibandingkan dengan metode modern seperti BERTopic.

Penelitian oleh (Apriyadi & Styawati, 2025) IndoBERT untuk menghasilkan representasi teks berbahasa Indonesia yang lebih kontekstual. Penelitian tersebut menunjukkan bahwa model berbasis Transformer memiliki kemampuan yang lebih baik dalam memahami hubungan semantik antar kata dibandingkan pendekatan tradisional. Namun, penelitian tersebut belum mengombinasikan IndoBERT dengan metode topic modeling, sehingga potensi representasi teks kontekstual untuk mendukung proses ekstraksi topik masih belum banyak dikaji. Celah penelitian ini menjadi salah satu dasar penerapan BERTopic berbasis IndoBERT dalam penelitian yang dilakukan.

Selanjutnya, penelitian oleh (Yan & Zhang, 2023) mengkaji penggunaan BERTopic dalam pemodelan topik pada data teks. Hasil penelitian menunjukkan bahwa BERTopic mampu menghasilkan topik yang lebih koheren dan relevan secara semantik dibandingkan dengan metode tradisional seperti LDA. Selain itu, penelitian ini juga menekankan pentingnya penggunaan metrik topic coherence dalam mengevaluasi kualitas topik yang dihasilkan.

Penelitian oleh (Ramadhan et al., 2025) yang membahas mengenai perbandingan model CNN dan IndoBERT dalam analisis sentimen berita politik Indonesia bertujuan untuk mengevaluasi performa kedua model dalam

mengklasifikasikan sentimen teks. Dalam pengembangan penelitian yang dilakukan, CNN berguna sebagai representasi salah satu model deep learning konvensional, sedangkan IndoBERT digunakan sebagai model berbasis transformer yang mampu memahami konteks bahasa secara lebih mendalam. Hasil penelitian menunjukkan bahwa IndoBERT mampu mencapai tingkat akurasi yang lebih tinggi dibandingkan dengan CNN dalam mengklasifikasikan sentimen berita politik.

Penelitian oleh (Dwiyono et al., 2024) Membandingkan performa beberapa model berbasis transformer, yaitu BERT, RoBERTa, dan IndoBERT dalam klasifikasi intent pada sistem chatbot. Tujuan penelitian adalah untuk mengetahui model yang paling efektif dalam memahami maksud pengguna berdasarkan input teks. Hasil penelitian menunjukkan bahwa IndoBERT mampu memberikan performa yang kompetitif dengan tingkat akurasi yang tinggi dalam mengklasifikasikan intent pengguna, serta menunjukkan kemampuan yang baik dalam memahami konteks bahasa Indonesia dibandingkan dengan model lainnya. Meskipun demikian, penelitian ini masih berfokus pada pendekatan supervised learning dan tidak mengkaji pemodelan topik atau eksplorasi pola teks secara unsupervised.

Dari hasil telaah literatur yang telah dilakukan, dapat disimpulkan bahwa model transformer seperti IndoBERT menunjukkan performa yang sangat baik dalam menangkap makna semantik dari teks berbahasa Indonesia. Di sisi lain, BERTopic sebagai metode topic modeling modern terbukti mampu menghasilkan topik yang lebih koheren dan mudah diinterpretasikan dibandingkan metode klasik

seperti LDA. Meskipun keduanya menawarkan keunggulan masing-masing, integrasi antara IndoBERT dan BERTopic masih jarang dieksplorasi dalam penelitian, khususnya pada data ulasan Google Maps di sektor kuliner lokal. Lebih lanjut, mayoritas penelitian terdahulu masih didominasi oleh pendekatan supervised learning untuk analisis sentimen, sementara pendekatan unsupervised untuk eksplorasi topik masih tergolong minim. Berangkat dari celah tersebut, penelitian ini bertujuan untuk mengkombinasikan IndoBERT dan BERTopic guna mengekstraksi tema-tema utama dari ulasan pengunjung warung kopi di Tulungagung.

2.2 Pemrosesan Bahasa Alami (*Natural Languange processing*)

Dalam ranah pemrosesan bahasa alami (NLP), langkah representasi teks menempati posisi yang sangat fundamental. Tahapan ini berfungsi sebagai jembatan utama yang menyambungkan data tekstual mentah dengan sistem komputasi. Pasalnya, perangkat komputer tidak memiliki kapasitas bawaan untuk memahami secara langsung struktur bahasa yang digunakan manusia sehari-hari. Oleh sebab itu, seluruh rangkaian input berupa kalimat, paragraf, ataupun dokumen harus diolah terlebih dahulu menjadi format numerik agar bisa diproses lebih lanjut oleh beragam metode analisis. Secara sederhana, langkah konversi ini kerap disebut dengan istilah vektorisasi atau embedding. Tujuan akhirnya adalah memetakan setiap unit kebahasaan ke dalam deretan angka yang tersusun dalam ruang berdimensi tertentu. Metode yang digunakan untuk mencapai tujuan tersebut terus berkembang seiring waktu, mulai dari yang paling elementer hingga yang berbasis kecerdasan buatan.

Bahasa alami memiliki karakteristik yang beragam dan kompleks, seperti adanya ambiguitas makna serta ketergantungan terhadap konteks penggunaan kata. Kondisi tersebut menyebabkan data bahasa tidak dapat diproses secara langsung oleh komputer tanpa melalui tahapan pengolahan tertentu. Oleh karena itu, NLP dikembangkan untuk mengubah bahasa manusia menjadi representasi yang lebih terstruktur sehingga dapat dipahami dan diolah oleh sistem komputasi dalam berbagai tugas analisis teks. (Alsayat & Elmitwally, 2020).

Sepanjang perkembangannya, Natural Language Processing (NLP) telah diadopsi secara ekstensif dalam beragam aplikasi teknologi, mulai dari analisis sentimen untuk mengukur persepsi masyarakat terhadap suatu produk atau isu, sistem rekomendasi yang mempersonalisasi pengalaman pengguna, hingga chatbot cerdas yang mampu berinteraksi layaknya manusia. NLP juga menjadi tulang punggung mesin pencari dan alat analisis media sosial yang berfungsi untuk memantau tren percakapan daring. Dalam ranah penelitian yang mengandalkan data tekstual, peran NLP menjadi semakin krusial. Teknik-teknik dalam NLP memungkinkan peneliti untuk mengekstraksi informasi bermakna dari data yang semula tidak beraturan seperti ulasan pelanggan, komentar media sosial, atau artikel berita dan akan diubah menjadi representasi terstruktur yang siap untuk dianalisis secara kuantitatif maupun kualitatif. Tanpa kontribusi NLP, data teks hanyalah sekumpulan simbol yang kehilangan konteks dan makna. Oleh karena itu, hampir dapat dikatakan bahwa NLP telah menjadi fondasi yang tak terpisahkan dari setiap penelitian analisis teks di era digital ini. Terutama dalam konteks pengolahan ulasan pengguna dan data opini, NLP menyediakan kerangka kerja yang memungkinkan

peneliti untuk menggali wawasan berharga dari data yang sebelumnya sulit diakses (Li, 2020).

Perkembangan NLP mengalami peningkatan yang signifikan dengan hadirnya pendekatan deep learning, terutama melalui arsitektur berbasis transformer seperti Transformer Model yang diperkenalkan oleh Google Research. Model ini mampu memahami hubungan antar kata dalam suatu kalimat secara kontekstual, sehingga menghasilkan representasi bahasa yang lebih akurat dibandingkan dengan metode tradisional seperti bag-of-words atau TF-IDF. Salah satu implementasi model transformer dalam bahasa Indonesia adalah IndoBERT yang dikembangkan dalam kerangka kerja IndoNLU oleh (Wilie, 2020). IndoBERT menggunakan pendekatan representasi kontekstual yang memungkinkan model memahami makna kata berdasarkan konteks kalimat, sehingga meningkatkan performa dalam berbagai tugas NLP seperti klasifikasi teks dan analisis sentimen.

2.3 Analisis Topik Pada Ulasan pengunjung

User Generated Content (UGC) merupakan konten yang dibuat dan dibagikan secara langsung oleh pengguna melalui platform digital, baik dalam bentuk teks, gambar, maupun video. UGC muncul sebagai bagian dari perkembangan ekosistem digital yang memungkinkan setiap pengguna untuk berpartisipasi aktif dalam menyampaikan pengalaman, pendapat, serta ulasan terhadap produk atau layanan yang mereka gunakan. Konten yang dihasilkan pengguna ini dinilai memiliki tingkat keaslian yang tinggi karena bersumber dari pengalaman nyata dan tidak sepenuhnya dikendalikan oleh pihak penyedia layanan. Salah satu bentuk UGC yang paling umum ditemukan adalah ulasan pengguna (user

review). Ulasan pengguna berisi penilaian, komentar, serta pengalaman pribadi konsumen setelah menggunakan suatu produk atau layanan. Ulasan ini sering dijadikan referensi oleh pengguna lain dalam pengambilan keputusan, sekaligus dimanfaatkan oleh penyedia layanan untuk mengevaluasi kualitas layanan yang diberikan. (Wayan et al., 2025).

Dalam konteks digital marketing dan komunikasi digital, UGC dipandang sebagai sumber informasi yang strategis. Konten yang dibuat oleh pengguna dianggap lebih dipercaya dibandingkan konten promosi resmi karena bersifat apa adanya dan mencerminkan pengalaman langsung. Penelitian terkait UGC menunjukkan bahwa ulasan dan komentar pengguna dapat memengaruhi kepercayaan, minat, serta persepsi pengguna lain terhadap suatu merek atau layanan, termasuk di platform media sosial seperti Instagram dan platform ulasan online (Rubyanti, 2020) (Costa & Aziz, 2021). Di Indonesia, praktik ulasan online berkembang seiring meningkatnya penetrasi internet dan penggunaan platform digital. Ulasan online banyak dijumpai pada sektor kuliner, pariwisata, aplikasi digital, hingga layanan publik. Melalui ulasan tersebut, konsumen dapat menyampaikan pengalaman positif maupun kendala yang mereka alami. Namun demikian, tidak semua ulasan bersifat objektif karena adanya potensi bias, ulasan palsu, maupun ulasan berbasis insentif yang dapat memengaruhi kualitas informasi yang diterima pengguna lain, Platform berbasis UGC berperan sebagai mediator yang memfasilitasi interaksi antara pengguna dan penyedia layanan. Platform ini memungkinkan pengguna untuk mempublikasikan ulasan secara terbuka, sementara penyedia layanan dapat memantau dan memahami persepsi pengguna

secara luas. Di sisi lain, tingginya volume UGC menimbulkan tantangan bagi platform, seperti perlunya moderasi konten, pengelolaan ulasan yang tidak relevan, serta perlindungan data pengguna. Hal ini menjadikan UGC sebagai sumber data yang kaya informasi, tetapi kompleks untuk dianalisis secara manual (Nidhal et al., n.d.)

Karakteristik utama ulasan pengguna sebagai UGC adalah bersifat tidak terstruktur dan menggunakan bahasa yang beragam. Banyak ulasan ditulis dengan bahasa informal, singkatan, bahasa daerah, serta kesalahan ejaan. Selain itu, satu ulasan sering kali membahas lebih dari satu aspek atau topik dalam satu teks. Kondisi ini menyebabkan proses analisis ulasan secara manual menjadi tidak efisien, terutama ketika jumlah data yang dianalisis sangat besar (Wayan et al., 2025) Beberapa penelitian menunjukkan bahwa UGC dalam bentuk ulasan dapat dimanfaatkan untuk menggali pola pembicaraan dan topik yang sering muncul di kalangan pengguna. Dengan mengelompokkan ulasan berdasarkan tema pembicaraan, organisasi dapat memahami isu utama, kebutuhan pengguna, serta aspek layanan yang paling sering mendapat perhatian. Pendekatan ini banyak digunakan dalam studi-studi analisis ulasan pengguna karena mampu memberikan gambaran yang lebih menyeluruh dibandingkan sekadar membaca ulasan satu per satu (Rangkuti et al., 2025)

Penelitian ini menjadikan ulasan pengguna sebagai sumber data primer untuk menggali tema-tema utama yang sering dibahas oleh para pengunjung berdasarkan pengalaman mereka. Ulasan yang termasuk dalam kategori UGC (User-Generated Content) ini dipilih karena kemampuannya dalam merefleksikan

opini dan persepsi pengguna secara spontan dan natural, sehingga diharapkan dapat memberikan wawasan yang lebih autentik dibandingkan data yang diperoleh melalui survei atau wawancara terstruktur. Melalui pemanfaatan UGC, penelitian ini bertujuan untuk mengidentifikasi aspek-aspek tertentu yang paling dominan muncul dalam diskusi pengguna, yang selanjutnya akan menjadi bahan baku dalam proses analisis topik. Dengan pendekatan ini, UGC tidak lagi dipandang sekadar sebagai konten ekspresif semata, tetapi bertransformasi menjadi instrumen analisis yang strategis untuk mengungkap pola-pola tematik dari data teks yang tidak terstruktur dan oleh karena itu diperlukan penelitian khusus (Rachman et al., 2021).

2.4 Analisis Topik Pada Data Teks

Salah satu tantangan utama dalam pengolahan data tekstual adalah bagaimana mengidentifikasi pola-pola tematik yang terkandung di dalamnya, terutama ketika jumlah dokumen sangat besar dan tidak terstruktur. Analisis topik hadir sebagai solusi untuk menjawab tantangan tersebut. Secara konseptual, metode ini berupaya untuk menemukan struktur tersembunyi dalam korpus teks dengan cara mengelompokkan dokumen berdasarkan kemiripan distribusi kata kunci yang digunakan. Hasil dari proses ini adalah serangkaian topik yang masing-masing merepresentasikan sekelompok dokumen dengan pembahasan serupa. Pendekatan ini telah banyak diterapkan di berbagai bidang, mulai dari riset pasar, analisis sentimen publik, hingga kajian literatur akademik, karena kemampuannya dalam menyederhanakan informasi kompleks menjadi tema-tema yang lebih mudah dipahami. Dengan pendekatan ini, peneliti dapat mengetahui topik apa saja yang

paling sering dibahas serta bagaimana kecenderungan pembahasan dalam suatu kumpulan data teks (Arifianto et al., 2020)

Analisis topik berbeda dengan analisis sentimen. Analisis sentimen berfokus pada penentuan sikap atau emosi pengguna, seperti positif, negatif, atau netral. Sementara itu, analisis topik tidak menilai perasaan pengguna, melainkan berfokus pada isi atau tema pembahasan yang muncul dalam teks. Dengan kata lain, analisis topik menjawab pertanyaan "apa yang dibicarakan", bukan "bagaimana perasaan penulis teks", Dalam penerapannya, analisis topik dapat dilakukan menggunakan berbagai pendekatan. Salah satu pendekatan yang umum digunakan adalah Latent Dirichlet Allocation (LDA), yang memodelkan dokumen sebagai kombinasi beberapa topik dan setiap topik sebagai kumpulan kata dengan probabilitas tertentu. Pendekatan ini banyak digunakan karena mampu mengidentifikasi topik dominan dari data teks berskala besar serta menentukan topik laten yang tersembunyi dalam kumpulan dokumen (Arifianto et al., 2020).

Selain pendekatan probabilistik, perkembangan pemrosesan bahasa alami juga mendorong penggunaan pendekatan berbasis representasi teks (embedding) dalam analisis topik. Pendekatan ini memanfaatkan representasi vektor teks untuk menangkap hubungan semantik antar kata secara kontekstual (Devlin et al., 2020). Analisis topik berbasis embedding dinilai lebih efektif dalam menangani teks pendek dan bahasa tidak baku karena mampu mengelompokkan dokumen berdasarkan kemiripan makna, bukan hanya kemunculan kata (Grootendorst, 2022).Berdasarkan karakteristik tersebut, analisis topik menjadi pendekatan yang relevan untuk diterapkan pada data ulasan pengguna. Analisis ini memungkinkan

peneliti mengidentifikasi aspek-aspek utama yang sering dibahas pengguna berdasarkan pengalaman mereka, sekaligus melihat kecenderungan dan pola pembahasan yang muncul.

2.5 Model bahasa Pra-latih indoBERT

Data teks memiliki sifat yang berbeda dengan data numerik atau kategorikal, sehingga memerlukan tahap transformasi khusus sebelum dapat dimasukkan ke dalam model komputasi. Transformasi ini bertujuan untuk merepresentasikan teks dalam ruang vektor (embedding) yang dapat diproses oleh algoritma machine learning. Seiring berkembangnya riset di bidang NLP, pendekatan berbasis model bahasa pra-latih menjadi semakin populer karena kemampuannya dalam menangkap makna kontekstual secara lebih baik dibandingkan metode statistik tradisional. Model-model ini pada dasarnya adalah jaringan saraf besar yang telah dilatih pada jutaan dokumen, sehingga memiliki pengetahuan kebahasaan yang luas. IndoBERT hadir sebagai salah satu model pra-latih yang disesuaikan untuk bahasa Indonesia. Dengan basis arsitektur Transformer, IndoBERT tidak hanya memahami makna kata secara leksikal, tetapi juga mempertimbangkan konteks kalimat secara utuh. Fleksibilitas IndoBERT tercermin dari kemampuannya untuk diadaptasikan pada berbagai tugas downstream dalam NLP, di antaranya analisis sentimen untuk mengukur opini publik, ekstraksi informasi untuk mengambil data terstruktur dari teks, serta pemodelan topik untuk mengetahui topik utama dalam kumpulan dokumen. Namun, daya tarik utama dari model ini bukan hanya pada fleksibilitasnya, melainkan juga pada efisiensinya. Karena IndoBERT telah melalui proses pra-pelatihan pada jutaan dokumen berbahasa Indonesia, model ini sudah

memiliki fondasi pemahaman bahasa yang kuat. Peneliti cukup melakukan fine-tuning pada dataset mereka sendiri—sebuah proses yang jauh lebih cepat dan murah dibandingkan harus membangun model dari awal. Dengan demikian, IndoBERT menawarkan keseimbangan ideal antara performa tinggi dan efisiensi.cnn

Bidirectional Encoder Representations from Transformers (BERT) merupakan salah satu contoh model bahasa berbasis pre-training yang sangat populer. Melalui pemanfaatan arsitektur Transformer yang mengintegrasikan mekanisme self-attention, model ini sanggup menganalisis keterkaitan antar-kata di dalam kalimat secara mendalam. Keunggulan utama BERT terletak pada kemampuannya memproses konteks dari dua arah—baik sebelum maupun sesudah kata—secara simultan. Pendekatan dua arah inilah yang membuatnya mampu menciptakan representasi teks yang jauh lebih kontekstual dan akurat ketimbang teknik tradisional.

Untuk mendukung pemrosesan teks berbahasa Indonesia, dikembangkan IndoBERT, yaitu model BERT yang dilatih menggunakan korpus bahasa Indonesia dalam jumlah besar. Model ini dirancang untuk memahami karakteristik bahasa Indonesia, termasuk penggunaan imbuhan, variasi struktur kalimat, serta ragam bahasa formal maupun informal yang banyak ditemukan pada teks digital. Dengan proses pelatihan tersebut, IndoBERT mampu menghasilkan representasi teks yang lebih sesuai dengan konteks bahasa Indonesia dibandingkan model BERT yang bersifat umum.

IndoBERT menghasilkan embedding kontekstual, yaitu representasi vektor yang mempertimbangkan keseluruhan konteks kalimat dalam proses

pembentukannya. Setiap token direpresentasikan dalam bentuk vektor numerik yang mengandung informasi semantik, sedangkan token khusus seperti [CLS] digunakan untuk mewakili makna umum dari suatu kalimat atau dokumen. Pendekatan ini berbeda dengan metode embedding konvensional, seperti Word2Vec dan GloVe, yang menghasilkan satu representasi tetap untuk setiap kata tanpa mempertimbangkan variasi konteks penggunaannya.

Pada penelitian ini, IndoBERT dimanfaatkan sebagai model representasi teks, bukan sebagai model klasifikasi. Embedding yang dihasilkan digunakan untuk merepresentasikan makna setiap ulasan pengguna sehingga hubungan semantik antar ulasan dapat ditangkap dengan lebih baik. Representasi tersebut kemudian menjadi dasar dalam proses ekstraksi topik menggunakan BERTopic, sehingga ulasan yang memiliki konteks serupa dapat dikelompokkan meskipun menggunakan pilihan kata yang berbeda. Dengan kemampuan tersebut, IndoBERT dinilai sesuai untuk mendukung proses analisis topik pada data ulasan pengguna yang umumnya bersifat singkat, beragam, dan tidak terstruktur (Baldwin, 2020).

2.6 BERTopic

BERTopic merupakan salah satu metode pemodelan topik berbasis embedding yang menggabungkan representasi teks kontekstual dengan teknik clustering untuk mengidentifikasi topik dalam kumpulan dokumen. Metode ini diperkenalkan oleh Maarten Grootendorst sebagai pendekatan modern yang mampu mengatasi keterbatasan metode tradisional seperti Latent Dirichlet Allocation (LDA) dalam memahami konteks semantik teks. BERTopic bekerja dengan memanfaatkan model bahasa pratelatih seperti IndoBERT untuk mengubah teks

menjadi representasi vektor (embedding). Representasi ini memungkinkan model untuk menangkap makna kata berdasarkan konteks kalimat secara menyeluruh.

Setelah proses embedding, dilakukan reduksi dimensi menggunakan algoritma Uniform Manifold Approximation and Projection (UMAP) untuk menyederhanakan ruang vektor berdimensi tinggi menjadi lebih rendah tanpa kehilangan informasi penting. Tahap selanjutnya adalah proses clustering menggunakan algoritma Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). Algoritma ini mengelompokkan dokumen berdasarkan kepadatan data dan mampu mengidentifikasi data yang dianggap sebagai noise. Pendekatan ini sangat efektif dalam menangani data teks yang tidak terstruktur dan memiliki variasi yang tinggi.

Proses pembentukan topik dalam BERTopic tidak berhenti pada tahap pengelompokan dokumen saja. Setiap klaster dokumen yang telah terbentuk masih perlu "diberi nama" atau representasi berupa kata-kata kunci yang mencerminkan inti pembahasan dari klaster tersebut. Di sinilah peran metode c-TF-IDF (Class-Based Term Frequency-Inverse Document Frequency) menjadi sangat krusial. Secara teknis, c-TF-IDF bekerja dengan menghitung seberapa sering sebuah kata muncul di dalam suatu topik tertentu, lalu membandingkannya dengan frekuensi kemunculan kata yang sama di seluruh topik lainnya. Kata-kata yang memiliki frekuensi tinggi secara eksklusif dalam satu topik akan memperoleh bobot yang tinggi dan dianggap sebagai kata kunci yang paling representatif. Keunggulan pendekatan ini tidak hanya terletak pada akurasi statistiknya, tetapi juga pada kemudahan interpretasinya. Berbeda dengan metode tradisional seperti LDA yang

sering menghasilkan topik yang ambigu atau tumpang tindih, c-TF-IDF pada BERTopic cenderung menghasilkan kata kunci yang lebih spesifik dan mudah dipahami. Hal ini disebabkan oleh mekanisme perhitungan yang mempertimbangkan distribusi kata secara per-topik, bukan hanya per-dokumen. Dengan demikian, BERTopic berhasil mengatasi salah satu kelemahan utama dari metode topic modeling konvensional, yaitu sulitnya memberikan interpretasi yang jelas terhadap topik yang dihasilkan.

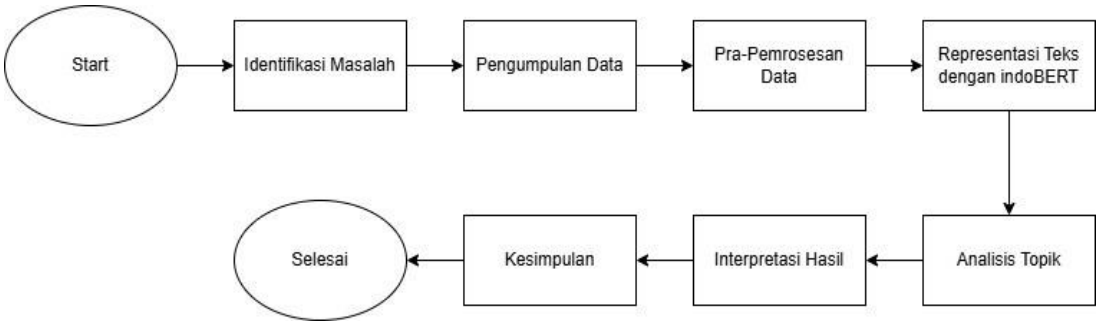
Dalam penelitian ini, BERTopic dipilih sebagai metode utama untuk mengidentifikasi topik-topik yang dominan dari ulasan pengguna warung kopi. Penggunaan IndoBERT sebagai model representasi teks diharapkan dapat semakin meningkatkan kualitas analisis, mengingat IndoBERT dirancang khusus untuk memahami nuansa bahasa Indonesia yang beragam—mulai dari bahasa formal hingga bahasa sehari-hari yang sering muncul dalam ulasan Google Maps. Dengan kombinasi BERTopic dan IndoBERT, penelitian ini menargetkan hasil analisis topik yang lebih akurat, koheren, dan kontekstual dibandingkan dengan pendekatan konvensional yang masih mengandalkan metode berbasis frekuensi kata.

BAB III

DESAIN DAN IMPLEMENTASI SISTEM

3.1 Desain Penelitian

Pada bagian ini akan dijelaskan langkah-langkah penelitian yang akan dilakukan penulis dalam menganalisis Topik Warung Kopi di Tulungagung. Desain penelitian akan disusun secara sistematis dan akan dijelaskan pada gambar 3.1



Gambar 3. 1 prosedur penelitian

Prosedur penelitian merupakan tahapan-tahapan yang dilakukan secara sistematis untuk mencapai tujuan penelitian. Pembuatan alur prosedur penelitian ini berfungsi sebagai gambaran yang jelas mengenai bagaimana alur kerja penelitian, sehingga setiap proses dapat dilakukan secara terstruktur dan terarah. Dengan adanya prosedur penelitian, diharapkan pelaksanaan penelitian dapat berjalan dengan lebih efektif dan sesuai dengan rancangan yang telah ditetapkan. Penelitian ini diawali dengan identifikasi permasalahan yang berkaitan dengan pemanfaatan ulasan pengguna sebagai sumber data teks. Setelah permasalahan ditetapkan, tahap selanjutnya adalah pengumpulan data berupa ulasan pengguna yang diperoleh dari platform digital yang relevan dengan objek penelitian. Data yang terkumpul kemudian diseleksi untuk memastikan kesesuaian dengan kebutuhan penelitian.

Tahap berikutnya adalah pra-pemrosesan data teks, yang bertujuan untuk membersihkan dan menyiapkan data agar siap dianalisis. Proses ini meliputi pembersihan teks dari karakter yang tidak diperlukan, normalisasi kata, serta penyesuaian format teks sehingga dapat mengurangi gangguan (noise) pada data. Data teks yang telah dipraproses kemudian direpresentasikan ke dalam bentuk vektor numerik menggunakan model bahasa pra-latih IndoBERT. Representasi teks ini digunakan untuk menangkap makna dan konteks dari setiap ulasan pengguna secara lebih baik. Embedding yang dihasilkan menjadi dasar dalam proses analisis topik. Tahap analisis topik dilakukan untuk mengidentifikasi tema-tema utama yang muncul dalam kumpulan ulasan pengguna berdasarkan kesamaan konteks pembahasan. Hasil analisis topik selanjutnya diinterpretasikan untuk memahami pola pembicaraan pengguna serta aspek-aspek yang paling sering dibahas. Tahap akhir penelitian adalah penarikan kesimpulan berdasarkan hasil analisis yang telah dilakukan.

3.1.1 Pengumpulan Data

Sumber data yang digunakan dalam penelitian ini berupa data teks ulasan pengguna yang diperoleh dari platform Google Maps, khususnya ulasan pengunjung terhadap salah satu warung kopi yang dijadikan objek penelitian. Ulasan pengguna dipilih sebagai sumber data karena memuat opini dan pengalaman pengunjung yang disampaikan secara langsung, sehingga dapat memberikan gambaran mengenai topik-topik pembahasan yang muncul berdasarkan interaksi pengguna dengan warung kopi tersebut, Pengumpulan data dilakukan dengan teknik pengambilan data secara otomatis (scraping) pada platform Google Maps.

Proses pengumpulan data dilakukan dengan mengambil ulasan pengguna yang tersedia secara publik tanpa melakukan interaksi langsung dengan pengguna. Data yang dikumpulkan berupa teks ulasan yang ditulis dalam bahasa Indonesia oleh pengunjung warung kopi yang menjadi objek penelitian.

Untuk memastikan kualitas data, dilakukan seleksi data berdasarkan beberapa kriteria. Data yang digunakan merupakan ulasan yang relevan dengan objek penelitian serta memiliki konten teks yang dapat dianalisis. Ulasan yang tidak mengandung teks yang bermakna, bersifat duplikat, atau tidak berkaitan dengan warung kopi yang diteliti tidak disertakan dalam proses analisis. Data yang telah dikumpulkan kemudian disimpan dalam format CSV untuk memudahkan proses pengolahan dan analisis pada tahap selanjutnya. Jumlah data yang digunakan dalam penelitian ini disesuaikan dengan kebutuhan analisis topik dan akan dijelaskan lebih lanjut pada Bab Hasil dan Pembahasan, sebagaimana ditunjukkan pada contoh data pada Tabel 3.1.

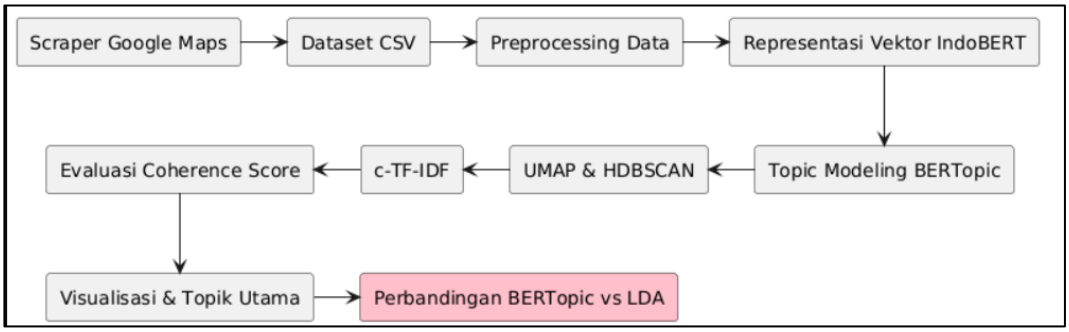
Tabel 3. 1 Pengumpulan Data

No	Nama Tempat	Link google Maps	rating	Name	Review text
1.	Warkop Waris ಬಾಡಿಮ್ಲ	https://maps.app.goo.gl/Fus8Ti4f4oG1dNuG9	5	James	Info kontaknya ka 🤷 saya dari Sumatra Utara Untuk lokasinya apa dekat dengan stasiun Tulungagung
2.	Warkop Waris ಬಾಡಿಮ್ಲ	https://maps.app.goo.gl/Fus8Ti4f4oG1dNuG9	5	Dwi Gimang	Tempat Nongkrong
3.	Warkop Waris ಬಾಡಿಮ್ಲ	https://maps.app.goo.gl/Fus8Ti4f4oG1dNuG9	5	Giri purbo wahyudi	wifi lancar poll, kursi banyak, kalo sore' suka stel lagunya S.I.D. cocok banget buat nyantai

No	Nama Tempat	Link google Maps	rating	Name	Review text
4.	Warkop Waris	https://maps.app.goo.gl/Fus8Ti4f4oG1dNuG9	5	Febri Nur biantoro	Tempat yang cocok untuk ngobrol dengan teman. Rekomen intinya.
5.	Warkop Waris	https://maps.app.goo.gl/Fus8Ti4f4oG1dNuG9	5	Moch. Arif Aji Seno	Jujur, ini UENAK,khas ! □□DE

3.1.1 Desain Sistem

Desain sistem penelitian ini dirancang untuk memberikan gambaran komprehensif mengenai alur proses analisis topik pada data ulasan pengguna menggunakan pipeline BERTopic dan model bahasa IndoBERT. Visualisasi desain sistem disajikan pada Gambar 3.2. Sistem diawali dengan scraper goole maps, yakni pengumpulan data ulasan pelanggan dari Google Maps pada warung kopi yang menjadi objek penelitian. Data yang telah terkumpul kemudian melalui tahap pra-pemrosesan (Preprocessing), meliputi cleaning, case folding, normalisasi (menggunakan kamus leksikon dan manual mapping), *Stopword Removal*, serta penyaringan dokumen pendek untuk memastikan kualitas data yang akan diolah. Data teks yang telah bersih kemudian direpresentasikan menjadi vektor numerik melalui tahap Representasi Vektor menggunakan model IndoBERT. Representasi ini menjadi dasar dalam Pemodelan Topik (Topic Modeling) yang melibatkan tiga proses utama: UMAP untuk reduksi dimensi, HDBSCAN untuk pengelompokan (clustering) ulasan dan c-TF-IDF untuk mencari kata kunci topik. Pada gambar 3.2 dibawah menunjukan alur bagaimana penelitian ini.



Gambar 3. 2 Desain Sistem

3.1.2 Preprocessing Data

Pra-pemrosesan data merupakan tahapan penting dalam pengolahan data teks sebelum dilakukan analisis lebih lanjut. Tahap ini bertujuan untuk membersihkan dan menyiapkan data teks agar dapat diproses secara optimal oleh metode pemrosesan bahasa alami. Data teks yang didapatkan dari ulasan pengguna pada platform digital memiliki data tidak terstruktur dan mengandung berbagai bentuk penulisan yang beragam, seperti penggunaan huruf kapital, simbol, emoji, serta kata tidak baku, proses pra-pemrosesan diperlukan untuk meningkatkan kualitas data dan mengurangi gangguan (noise) pada teks (Jakhotiya et al., 2022).

Sebelum data teks dapat dianalisis lebih lanjut, diperlukan serangkaian proses pembersihan dan penyeragaman yang dikenal dengan istilah pra-pemrosesan data (preprocessing). Langkah ini bertujuan untuk menyaring informasi yang benar-benar relevan dan sekaligus membuang elemen-elemen yang tidak diperlukan seperti tanda baca berlebihan, angka, atau kata-kata yang tidak bermakna (stopwords). Dalam konteks analisis topik, kualitas hasil akhir sangat ditentukan oleh kualitas data yang menjadi input. Data yang dipra-proses secara optimal akan menghasilkan vektor representasi teks yang lebih akurat, sehingga proses

pengelompokan dokumen ke dalam topik-topik tertentu menjadi lebih efektif dan menghasilkan interpretasi yang lebih bermakna. Menyadari pentingnya tahapan ini, penelitian merancang alur pra-pemrosesan yang disesuaikan dengan dua pertimbangan utama. Pertama, karakteristik data ulasan Google Maps yang cenderung bersifat informal, menggunakan bahasa sehari-hari, serta mengandung variasi ejaan dan singkatan. Kedua, kebutuhan analisis topik berbasis representasi teks yang akan diproses menggunakan model IndoBERT, yang memerlukan masukan teks dalam format tertentu agar dapat menghasilkan representasi semantik yang optimal.

1. *Cleaning*

Cleaning merupakan pembersihan data yang merupakan proses awal dari text-processing yang bertujuan untuk menghilangkan karakter, simbol maupun elemen teks yang tidak relevan dengan analisis, proses ini sangat penting untuk dilakukan karena data hasil kusioner terbuka yang dikumpulkan umumnya mengandung unsur-unsur seperti tanda baca, angka ataupun karakter khusus yang dapat menganggu proses pembobotan dan klasifikasi. Hasil dari proses tersebut ditunjukkan pada Tabel 3.2.

Tabel 3. 2 Proses Cleaning

Sebelum *cleaning*	Kopinya ENAK banget!!!! 😍 Tempatnya nyaman & adem
Sedudah *cleaning*	Kopinya ENAK banget Tempatnya nyaman adem

2. *Case Folding*

Case folding merupakan proses normalisasi huruf pada teks agar sekuruh karakter berada dalam keadaan sama yaitu dalam bentuk huruf kecil. Proses ini

dilakukan untuk menghilangkan makna antara huruf kecil dengan huruf besar yang pada dasarnya memiliki konteks yang sama. Hasil dari proses tersebut ditunjukkan pada Tabel 3.3.

Tabel 3. 3 Proses Case Folding

Sebelum *Case Folding*	Kopinya ENAK banget Tempatnya nyaman adem.
Setelah *Case Folding*	kopinya enak banget tempatnya nyaman adem

3. Normalisasi Kata

Normalisasi kata merupakan proses penyesuaian kata-kata tidak baku, singkatan, atau bentuk penulisan informal menjadi kata baku yang sesuai dengan kaidah bahasa. Tahap normalisasi kata akan dilakukan untuk memetakan kata tidak baku atau bahasa gaul (slang) ke dalam bentuk baku atau sesuai dengan KBBI. Dalam proses normalisasi, didukung oleh dataset, yaitu Colloquial Indonesian Lexicon, yang berisi mengenai pemetaan ribuan kata yang tidak baku ke dalam bentuk bakunya. Selain itu, juga dilakukan manual mapping tambahan untuk mengakomodasi istilah arti dari bahasa Jawa Timur seperti nyethe atau kopine yang belum tercover dalam kamus Colloquial Indonesian Lexicon. Proses ini sangat krusial untuk memastikan bahwa model dari IndoBERT tidak salah dalam menginterpretasikan variasi penulisan kata yang memiliki makna kontekstual yang sama. ditunjukkan pada Tabel 3.4.

Tabel 3. 4 Proses Normalisasi Kata

Sebelum Normalisasi Kata	kopinya enak bgt tmptnya nyaman adem
Sesudah Normalisasi Kata	kopinya enak banget tempatnya nyaman adem

4. *Stopword Removal*

Stopword Removal merupakan proses penghapusan kata-kata umum (stopwords) yang sering muncul dalam teks tetapi tidak memiliki makna signifikan dalam proses analisis. Kata-kata seperti “dan”, “yang”, “adalah”, “dengan”, “saya”, atau “itu” biasanya muncul dalam hampir semua kalimat, namun tidak memberikan kontribusi bermakna terhadap pengenalan pola sentimen. Oleh karena itu, kata-kata tersebut dihapus agar analisis dapat menyaring pada kata-kata penting yang memiliki nilai informasi tinggi. Dalam penelitian ini, proses *Stopword Removal* dilakukan setelah tahap *Tokenizing* menggunakan daftar kata berhenti (stopword list) berbahasa Indonesia yang disediakan oleh pustaka pemrosesan teks pada bahasa pemrograman Python, seperti Natural Language Toolkit (NLTK). Sistem akan secara otomatis memeriksa setiap token dan menghapus kata yang termasuk dalam daftar stopwords. Hasil dari proses tersebut ditunjukkan pada Tabel 3.5

Tabel 3. 5 proses *Stopword Removal*

Sebelum *Stopword Removal*	kopinya enak banget dan tempatnya nyaman di warung ini
Sesudah *Stopword Removal*	kopinya enak banget tempatnya nyaman warung ini

5. *Tokenizing*

Tokenizing adalah sebuahanal proses pemecahan teks menjadi unit-unit yang lebih kecil yang disebut token, seperti kata atau subkata, agar dapat diproses oleh model pemrosesan bahasa alami. Dalam penelitian ini, tokenisasi tidak dilakukan secara manual pada tahap pra-pemrosesan data. Proses tokenisasi dilakukan secara otomatis oleh model bahasa pra-latih IndoBERT menggunakan mekanisme sub-word tokenization. Dengan mekanisme ini, teks dipecah menjadi unit subkata yang sesuai

dengan kosakata model, sehingga kata yang tidak terdapat secara utuh dalam kosakata tetap dapat diproses dengan baik. Hasil dari proses tersebut ditunjukkan pada Tabel 3.6.

Tabel 3. 6 Proses *Tokenizing*

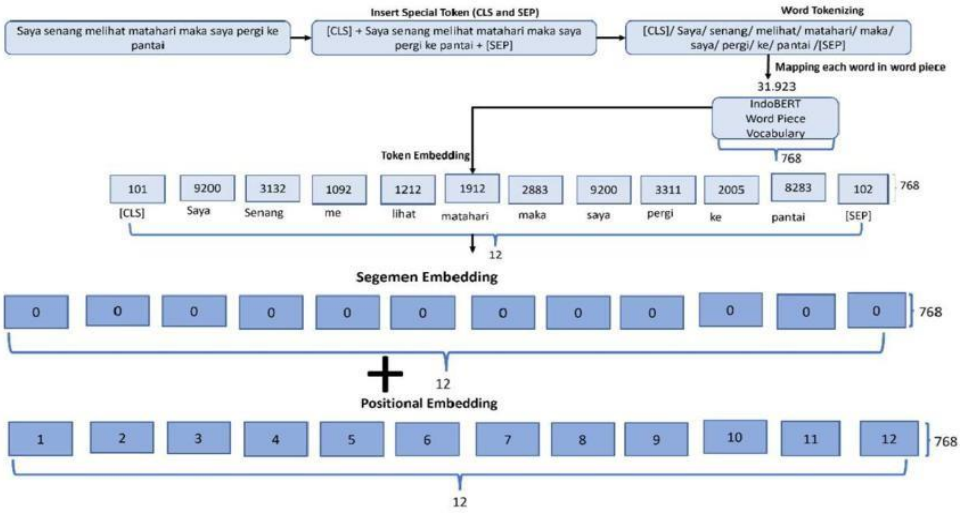
Sebelum *Tokenizing*	mungkin kalau dampak positif bisa menjadi ruang diskusi bagi akademisi meningkatkan pendapatan masyarakat yang berjualan cethe menjadi budaya
Sesudah *Tokenizing*	[“dampak”, “positif”, “ruang”, “diskusi”, “akademi”, “tingkat”, “pendapat”, “masyarakat”, “jual”, “cethe”, “budaya”]

3.1.3 Representasi Teks indoBERT

Setelah melalui tahap pra-pemrosesan data, teks ulasan pengguna direpresentasikan ke dalam bentuk numerik menggunakan model bahasa pra-latih IndoBERT. Representasi teks diperlukan agar data teks dapat diproses secara komputasional dan digunakan pada tahap analisis topik. IndoBERT dipilih karena mampu menghasilkan embedding kontekstual yang sesuai dengan karakteristik Bahasa Indonesia. Proses representasi teks dilakukan dengan memasukkan teks hasil pra-pemrosesan ke dalam tokenizer IndoBERT. Tokenisasi dilakukan secara otomatis menggunakan mekanisme WordPiece, di mana setiap teks ditambahkan token khusus CLS di awal kalimat dan SEP di akhir kalimat. Mekanisme ini memungkinkan teks dipecah menjadi unit sub-kata sehingga kata yang jarang muncul tetap dapat direpresentasikan dengan baik.

Selanjutnya, setiap token diubah menjadi token embedding dan digabungkan dengan positional embedding untuk mempertahankan informasi urutan kata dalam kalimat. Proses ini menghasilkan representasi vektor berdimensi

tetap yang merepresentasikan makna dan konteks dari setiap ulasan pengguna secara menyeluruh. Pada penelitian ini, model IndoBERT digunakan hanya sebagai model representasi teks, tanpa dilakukan proses pelatihan ulang, fine-tuning, maupun klasifikasi. Embedding yang dihasilkan kemudian digunakan sebagai input pada tahap analisis topik untuk mengidentifikasi pola dan tema utama yang muncul dalam ulasan pengguna. Gambar 3.3 contoh proses Representasi Teks dengan IndoBERT.



Gambar 3. 3 Proses indoBERT

3.1.4 Proses Analisis Topik

Proses analisis topik pada penelitian ini dilakukan dengan mengimplementasikan keranga kerja (framework) yaitu BERTopic, merupakan sebuah metode pemodelan topik modern yang bekerja memanfaatkan representasi embedding teks dari model pra-latih indoBERT. indoBERT memungkinkan sistem untuk mengenali maupun mengidentifikasi tema-tema utama dalam kumpulan ulasan pelanggan dengan tingkat akurasi semantik yang tinggi , karena setiap dokumen ulasan akan dipetakan kedalam numerik vektor yang berdimensi 768 yang

bisa menangkap makna kontekstual secara profesional. Vektor embedding juga berfungsi sebagai pondasi utama dalam mengukur kedekatan antar konteks ulasan atau kata tanpa kehilangan informasi linguistik yang krusial.

Dalam alur kerja BERTopic, data yang telah masuk ke dalam ruang vektor berdimensi tinggi tidak akan langsung dikelompokkan, melainkan akan melewati tahapan reduksi dimensi menggunakan algoritma UMAP (Uniform Manifold Approximation and Projection). Tahapan UMAP sendiri memiliki fungsi, yaitu untuk menyederhanakan representasi data agar proses komputasi lebih efisien dan struktur klaster akan menjadi lebih padat (Grootendorst, 2020) . Setelah dimensi data optimal, proses pengelompokan dilakukan dengan HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) yang berfungsi sebagai awal untuk membentuk klaster dokumen berdasarkan kemiripan semantiknya secara otomatis tanpa perlu menentukan jumlah topik di awal.

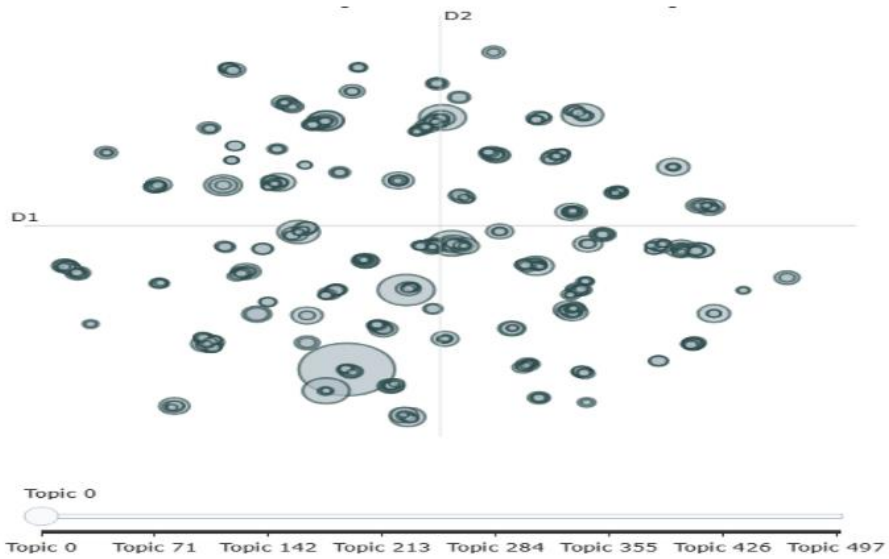
Sistem akan memasuki tahap ekstraksi kata kunci menggunakan metode c-TF-IDF (class-based Term Frequency-Inverse Document Frequency). Metode ini secara spesifik akan menghitung bobot kepentingan kata dalam setiap klaster untuk menghasilkan kata-kata yang representatif bagi masing-masing topik yang sudah ditemukan. Keseluruhan dari rangkaian ini akan diakhiri dengan Topic Coherence yang akan memastikan bahwa setiap topik memiliki makna yang baik. Pembahasan lebih mendalam mengenai setiap komponen teknis dalam pipeline analisis topik dipaparkan pada subbagian berikut:

1. BERTopic

BERTopic merupakan metode topic modeling modern yang memanfaatkan representasi teks berbasis transformer yang menghasilkan topik lebih koheren dan mudan untuk diinterpretasikan, menurut (Hutama & Suhartono, 2022)BERTopic merupakan teknik topic modeling yang memanfaatkan representasi teks berbasis BERT dan C-TF-IDF untuk menghasilkan topik dalam kumpulan dokumen. Dalam prosesnya, BERTopic bekerja melalui tahapan utama yang saling berhubungan. Tahap pertama merupakan pembentukan document embedding, yaitu mengubah setiap dokumen teks menjadi representasi vektor numerik. Representasi vektor numerik sendiri memungkinkan sistem menangkap makna kontekstual dari kata dalam sebuah kalimat, sehingga dokumen yang memiliki arti yang hampir sama akan memiliki jarak vektor yang lebih dekat dengan ruang representasi numerik.

Menurut penelitian yang dilakukan oleh (Wahyuni et al., 2025) Yaitu BERTopic digunakan untuk membahas opini publik terhadap program makanan bergizi gratis (MBG) melalui komentar youtube. Dalam penelitian ini, data yang dianalisis adalah 19.843 komentar yang berasal dari 71 video youtube dan dianalisis, melewati tahapan proses processing, document embedding, reduksi dimensi menggunakan UMAP, clustering menggunakan HDBSCAN, serta representasi topik menggunakan metode C-TF-IDF. Hasil analisis dengan BERTopic mampu menunjukkan bahwa model BERTopic mampu mengelompokkan komentar publik menjadi beberapa topik utama yang menggambarkan beragamnya persepsi masyarakat terhadap program tersebut.

Penelitian lain yang dilakukan oleh (Wesly M Sihombing, 2025) Juga memanfaatkan BERTopic untuk menganalisis isu kesehatan publik di media sosial X dengan mengombinasikan metode IndoBERTweet untuk analisis sentimen dan BERTopic untuk ekstraksi topik. Penelitian tersebut menggunakan 6.740 tweet sebagai dataset dan menghasilkan 44 topik diskusi, dengan sembilan topik utama yang paling dominan dalam percakapan publik mengenai isu kesehatan seperti kesehatan mental, vaksinasi, dan penyakit menular. Hasil evaluasi menunjukkan bahwa model BERTopic menghasilkan nilai coherence score sebesar 0,7154 dan diversity score sebesar 0,7139, yang menunjukkan bahwa topik yang dihasilkan memiliki keterkaitan semantik yang baik serta keragaman topik yang cukup tinggi.



Gambar 3. 4 Intertopic Distance

Secara umum, BERTopic bekerja melalui beberapa tahapan yang saling berkaitan untuk menghasilkan topik dari kumpulan dokumen teks. Tahap awal dimulai dengan proses document embedding, yaitu mengubah setiap dokumen menjadi representasi vektor menggunakan model bahasa berbasis Transformer,

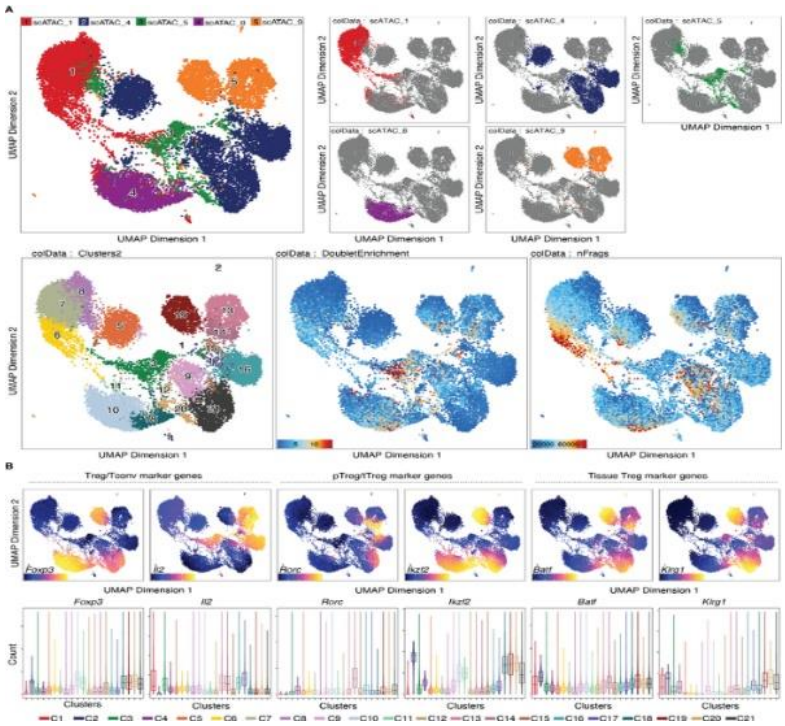
seperti BERT atau Sentence-BERT. Representasi tersebut memungkinkan sistem memahami makna dan hubungan kontekstual yang terdapat pada setiap dokumen. Tahap berikutnya adalah reduksi dimensi menggunakan Uniform Manifold Approximation and Projection (UMAP). Proses ini bertujuan untuk menyederhanakan embedding berdimensi tinggi menjadi representasi yang lebih ringkas tanpa menghilangkan informasi penting, sehingga memudahkan proses pengelompokan dokumen.

Selanjutnya, dokumen yang telah direduksi dimensinya dikelompokkan menggunakan algoritma clustering untuk membentuk kumpulan dokumen dengan karakteristik yang serupa. Setelah proses clustering selesai, BERTopic menerapkan pembobotan kata menggunakan c-TF-IDF untuk mengidentifikasi kata-kata yang paling representatif pada setiap topik yang terbentuk. Melalui kombinasi representasi teks berbasis embedding, reduksi dimensi, clustering, dan pembobotan kata, BERTopic mampu menghasilkan topik yang memiliki keterkaitan semantik yang baik. Pendekatan ini memungkinkan model menangkap hubungan makna antar dokumen secara lebih efektif dibandingkan metode topic modeling tradisional yang hanya mengandalkan frekuensi kemunculan kata.

2. Reduksi Dimensi dengan UMAP

Uniform Manifold Approximation and Projection (UMAP) Merupakan algoritma reduksi dimensi yang berfungsi memproyeksikan data berdimensi tinggi ke dalam ruang berdimensi yang lebih rendah dengan tetap mempertahankan struktur topologi datanya. UMAP bekerja dengan asumsi bahwa data berdimensi tinggi berada pada suatu manifold berdimensi rendah sehingga dapat

direpresentasikan kembali pada ruang dimensi yang lebih kecil tanpa kehilangan informasi penting dalam data tersebut (Mcinnes et al., 2020).



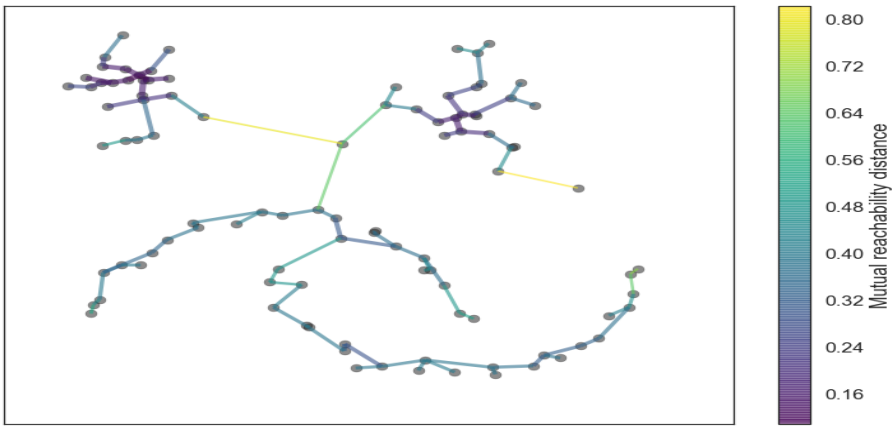
Gambar 3. 5 Reduksi UMAP

Setiap data atau dokumen dalam dataset direpresentasikan sebagai titik tunggal pada visualisasi UMAP. Tingkat kemiripan karakteristik antar-data diindikasikan oleh posisi titik tersebut, di mana kedekatan jarak menunjukkan hubungan atau kesamaan yang lebih intens dibandingkan dengan posisi yang saling menjauh. Di samping itu, pengelompokan (cluster) yang terbentuk setelah reduksi dimensi ditandai oleh perbedaan warna pada masing-masing titik. Melalui pemetaan dari ruang berdimensi tinggi ke rendah ini, tampak bahwa UMAP berhasil mempertahankan struktur lokal data. Dampaknya, dokumen-dokumen yang memiliki kemiripan tetap berada di area yang sama tanpa kehilangan detail

informasi penting mengenai interaksi antar-data. Kemampuan inilah yang membuat UMAP sangat andal dalam mereduksi kompleksitas data guna mendukung tahapan analisis clustering berikutnya.

3. HDBSCAN

Sebagai bentuk penyempurnaan dari metode DBSCAN, HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) hadir sebagai salah satu algoritma pengelompokan berbasis densitas. Keunggulan utama dari metode ini adalah kemampuannya dalam mengidentifikasi berbagai kelompok data (cluster) yang memiliki tingkat kepadatan berbeda tanpa menuntut pengguna menentukan jumlah klaster di awal secara manual. mekanisme kerja algoritma ini bertumpu pada estimasi kerapatan data untuk menyusun sebuah struktur klaster yang bersifat hierarkis (Campello et al., 2013),



Gambar 3. 6 HDBSCAN

Pada metode clustering berbasis kepadatan, suatu cluster didefinisikan sebagai kumpulan titik data yang memiliki kepadatan lebih tinggi dibandingkan dengan area di sekitarnya. Titik data yang tidak memenuhi kriteria kepadatan tertentu akan dikategorikan sebagai noise atau outlier. Pendekatan ini

memungkinkan algoritma untuk mengidentifikasi bentuk klaster yang kompleks serta menangani data yang mengandung noise secara efektif. Dalam proses kerjanya, HDBSCAN mengembangkan konsep DBSCAN dengan membangun hierarki klaster menggunakan estimasi kepadatan sehingga dapat mendeteksi klaster dengan kepadatan yang berbeda. Pendekatan ini menghasilkan struktur klaster bertingkat yang kemudian disederhanakan untuk memperoleh klaster yang paling signifikan berdasarkan ukuran stabilitas klaster.

4. c-TF-IDF

Penerapan metode Class-Term Frequency-Inverse Document Frequency (c-TF-IDF) dalam studi ini ditujukan untuk menjaring kata kunci representatif dari masing-masing topik yang dibentuk oleh model BERTopic. Sebagai variasi perluasan dari TF-IDF konvensional—yang jamak digunakan untuk menilai signifikansi kata dalam suatu berkas teks—c-TF-IDF menawarkan pendekatan berbeda. Karakteristik pembeda utamanya adalah kalkulasi pembobotan kata yang dilakukan pada level cluster atau kelas dokumen. Melalui teknik ini, seluruh dokumen yang tergabung dalam satu klaster diakumulasikan menjadi satu dokumen perwakilan topik, sehingga kata-kata yang paling khas dan relevan pada topik tersebut dapat ditonjolkan dengan baik (Grootendorst, 2020).

Metode C-TF-IDF pada dasarnya menggabungkan dua indikator utama, yaitu frekuensi kata dalam satu klaster dan frekuensi kebalikan antar-klaster yang dimodifikasi dari IDF. Yang pertama, atau kerap disebut TF, digunakan untuk menilai intensitas kemunculan suatu istilah pada suatu kumpulan dokumen yang telah dikelompokkan. Jika sebuah kata muncul berulang kali di dalam satu klaster,

itu menandakan bahwa kata tersebut memiliki relevansi yang kuat terhadap karakteristik klaster tersebut. Adapun formula untuk mengukur indikator pertama ini telah dirumuskan secara jelas pada persamaan 3.1(Grootendorst, 2020).

$$tf_{t,c} = \frac{ft,c}{\sum ft,c} \quad 3.1$$

Keterangan:
$tf_{t,c}$ = Nilai Term Frequency dari kata t pada kelas atau topik c
ft,c = Jumlah kemunculan kata t pada kelas c
$\sum ft,c$ = total seluruh kata dalam kelas c

Setelah nilai TF diperoleh, selanjutnya menghitung dari ICF yang menunjukkan seberapa penting seluruh kata dibandingkan dengan seluruh kelas atau topik dalam kumpulan dokumen. Kata yang muncul di banyak topik akan memiliki nilai yang lebih kecil karena dianggap kurang spesifik dalam mempresentasikan suatu topik. Rumus perhitungan ICF ditunujukan pada persamaan 3.2

$$icf_t = \log\left(1 + \frac{A}{tft}\right) \quad 3.2$$

Keterangan:
$icft$ = nilai inverse class frequency untuk kata t
A = rata-rata jumlah kata pada seluruh kelas
tf_t = frekuensi kemunculan t pada seluruh kelas

Tahap selanjutnya setelah nilai ICF didapatkan yaitu menghitung nilai dari C-TF-IDF yang merupakan kombinasi antara nilai TF dengan ICF. Nilai ini akan digunakan untuk menentukan tingkat kepentingan suatu kata dalam merepresentasikan topik tertentu. Rumus perhitungan C-TF-IDF ditunjukan pada persamaan 3.3

$$W_{t,c} = tf_{t,c} \times log\left(1 + \frac{A}{tf_t}\right) \quad 3.3$$

Keterangan:
Wt_c = bobot c-TF-IDF dari kata t pada topik c
$tf_{t,c}$ = nilai TF dari kata t pada kelas c
A = rata-rata jumlah kata pada seluruh kelas
tf_t = frekuensi kata t pada seluruh kelas

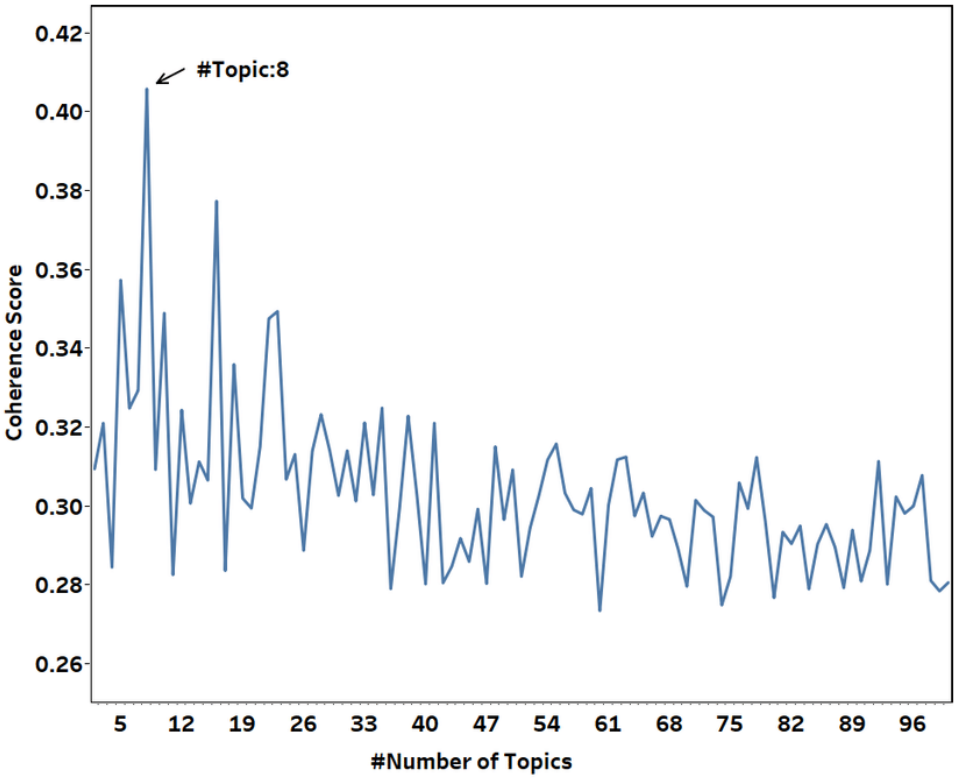
Hasil dari proses ini adalah representasi kata kunci dari setiap topik, kata-kata dengan bobot C-TF-IDF tertinggi akan dianggap sebagai kata kunci paling merepresentasikan topik tersebut, Representasi ini akan digunakan untuk menginterpretasikan setiap topik yang dihasilkan oleh model BERTopic.

3.1.5 Topic coherence

Topic coherence merupakan metrik evaluasi yang umum diaplikasikan guna mengukur tingkat kualitas dari topik yang diekstraksi lewat metode pemodelan topik (*topic modelling*). Kegunaan utama dari metrik ini adalah untuk mengukur kedekatan hubungan semantik antar-kata dalam sebuah topik, serta memastikan apakah topik tersebut dapat diinterpretasikan dengan mudah oleh manusia. Peningkatan skor *topic coherence* mengindikasikan kualitas topik yang kian optimal, sebab kosakata yang menyusun topik tersebut memiliki keterikatan makna yang solid satu sama lain.

Dalam penelitian sebelumnya, topic coherence sendiri digunakan sebagai indikator untuk membandingkan performa berbagai metode topic modelling seperti LDA, NMF dan BERTopic. Hasil dari penelitian tersebut menunjukkan bahwa model dengan nilai coherence yang lebih tinggi cenderung menghasilkan topik yang lebih konsisten dan interpretatif (Wahyana et al., 2024). Pada tataran konseptual, topic coherence berfungsi sebagai indikator untuk menilai sejauh mana kata-kata yang tergabung dalam sebuah topik saling berhubungan secara semantik.

Pengukuran ini didasarkan pada pola kehadiran kata-kata tersebut secara simultan di dalam korpus dokumen yang sama.



Gambar 3. 7 Topic Coherence

Studi ini menerapkan metrik *c_v* sebagai instrumen dalam mengalkulasi nilai koherensi topik. Metrik tersebut mengintegrasikan NPMI guna menganalisis relasi probabilitas antar-kata, yang kemudian diteruskan dengan kalkulasi *cosine similarity* untuk menilai kedekatan semantik antar-vektor kata pada keseluruhan topik. Penggabungan kedua teknik ini mampu menyajikan hasil penilaian yang paling selaras dengan penalaran atau persepsi manusia jika dikomparasikan dengan metrik evaluasi konvensional lainnya.

3.2 Implementasi Penelitian

Bagian ini memaparkan eksekusi riil dari desain penelitian yang telah direncanakan sebelumnya. Proses implementasi dilakukan guna mewujudkan tiap alur kerja, yang diawali dengan akuisisi data, prapemrosesan (preprocessing), penyusunan representasi teks berbasis IndoBERT, dilanjutkan ke tahap penarikan topik lewat BERTopic, hingga diakhiri dengan pengujian topic coherence serta komparasi terhadap metode LDA. Eksekusi seluruh rangkaian sistem ini dibangun berbasis bahasa pemrograman Python pada platform Google Colaboratory, dengan mengintegrasikan sejumlah library yang relevan untuk komputasi dan analisis data tekstual.

1. Tahap Cleaning

Tahapan cleaning merupakan proses awal preprocessing yang betujuan untuk menghapus URL dan karakter selain huruf, angka, spasi dan underscore yang akan dibersihkan pada data teks. Pada tahap ini digunakan fungsi re.sub() dari regular expression pada python.

```
# 2 CLEANING
    clean_text = re.sub(r'http\S+|www\S+|https\S+', '', case_text)
    clean_text = re.sub(r'[^a-z0-9\s_]', ' ', clean_text)
```

Gambar 3. 8 Sourcode Tahap Cleaning

Pada Gambar 3.8 ditunjukkan implementasi kode pada tahap cleaning data teks. Proses ini diawali dengan pemanfaatan modul re untuk melakukan pembersihan teks. Pada proses cleaning data yang dilakukan, tidak berfokus pada mention dan hashtag karena jarang ditemukan pada ulasan Google Maps.

2. Case Folding

Proses *case folding* diaplikasikan dengan tujuan menyamakan format seluruh rangkaian kata menjadi huruf kecil (*lowercase*). Langkah prapemrosesan ini krusial guna mengeliminasi inkonsistensi data yang dipicu oleh variasi penulisan huruf kapital. Pada penelitian ini, eksekusi *case folding* diimplementasikan dengan memanfaatkan fungsi bawaan .lower() yang tersedia pada Python. Sebagai ilustrasi, frasa "Kopi Enak" secara otomatis akan dikonversi menjadi "kopi enak".

```
# 1 CASE FOLDING
    case_text = text.lower()
    print("\nCASE FOLDING")
    print("Sebelum:", text)
    print("Sesudah:", case_text)
```

Gambar 3. 9 Sourecode *Case Folding*

Pada Gambar 3.9 ditampilkan kode pada tahap case folding yang berfungsi mengubah seluruh kata ke dalam huruf kecil. Hasil dari case folding disimpan pada kolom baru dan dimanfaatkan pada tahap preprocessing selanjutnya.

3. Normalisasi Kata

Tahapan normalisasi kata dijalankan dengan tujuan mengonversi kosakata nonbaku maupun istilah gaul (slang) ke dalam format standar sesuai kaidah bahasa. Langkah ini krusial guna menjaga konsistensi data selama tahapan analisis berlangsung, mengingat data berupa ulasan (review) umumnya didominasi oleh penggunaan bahasa informal.

```
import pandas as pd
import re
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory


# load lexicon
lexicon = pd.read_csv("/content/colloquial-indonesian-lexicon.csv")

    normalized_words = [norm_dict.get(w, w) for w in tokens]

    print("\nNORMALISASI")
    print("Sebelum:", tokens)
    print("Sesudah:", normalized_words)
```

Gambar 3. 10 Source code Normalisasi kata

Rangkaian proses normalisasi dalam studi ini memanfaatkan kamus colloquial Indonesian lexicon, yang memuat daftar padanan antara kosakata slang dengan bentuk bakunya. Berkas kamus tersebut awalnya disimpan dalam format CSV, untuk kemudian ditransformasikan ke dalam struktur data dictionary menggunakan Python.

4. *Stopword Removal*

Stopword Removal merupakan prosedur eliminasi kata-kata umum yang dinilai kurang memberikan signifikansi informasi dalam proses analisis, seperti konjungsi atau preposisi "dan", "yang", serta "di". Dalam penelitian ini, skema pengeliminasian kata tersebut mengacu pada daftar standar yang telah dikonfigurasi sebelumnya. Kosakata yang teridentifikasi dalam daftar tersebut bakal dibersihkan dari dokumen agar tidak mendistorsi atau memengaruhi kualitas ekstraksi topik pada tahapan berikutnya.

```
# 5 STOPWORD REMOVAL
filtered_words = [w for w in normalized_words if w not in stopwords]

print("\nSTOPWORD REMOVAL")
print("Sebelum:", normalized_words)
print("Sesudah:", filtered_words)

final_text = " ".join(filtered_words)

print("\nFINAL TEXT:", final_text)

return final_text
```

Gambar 3. 11Sourcode *Stopword Removal*

Gambar 3.11 menampilkan implementasi kode pada tahap stopword removal, yaitu proses yang bertujuan untuk mengurangi kata-kata umum yang memiliki kontribusi rendah terhadap proses analisis. Pada penelitian ini, proses tersebut dilakukan dengan memanfaatkan daftar stopword bahasa Indonesia yang disediakan oleh pustaka Sastrawi. Hasil tokenisasi dan normalisasi kemudian diperiksa satu per satu untuk mempertahankan kata-kata yang tidak termasuk dalam daftar stopword. Mekanisme yang digunakan berupa proses filtering, di mana kata-kata yang dianggap kurang informatif akan dihapus dari data teks. Melalui tahapan ini, data yang dihasilkan menjadi lebih relevan dan berfokus pada kata-kata penting, sehingga dapat mendukung proses representasi teks dan ekstraksi topik pada tahap berikutnya.

5. Tokenizing

Tokenizing merupakan proses pemecahan teks menjadi unit-unit kata untuk memudahkan analisis. Pada tahap pra-pemrosesan, tokenisasi awal dilakukan menggunakan metode sederhana berbasis spasi dengan fungsi .split() pada Python.

Proses ini bertujuan untuk kebutuhan pembersihan teks, seperti pada tahap *Stopword Removal*.

Namun demikian, pada tahap representasi teks, proses tokenisasi utama dilakukan secara otomatis oleh tokenizer IndoBERT menggunakan mekanisme WordPiece. Dengan demikian, tokenisasi dalam penelitian ini dilakukan dalam dua tahap, yaitu tokenisasi sederhana pada preprocessing dan tokenisasi lanjutan pada saat proses embedding menggunakan IndoBERT. Sebagai contoh, teks "kopi enak banget" akan diubah menjadi token ["kopi", "enak", "banget"] pada tahap tokenisasi awal.

```
# 3 TOKENIZING
tokens = clean_text.split()

print("\nTOKENIZING")
print("Sebelum:", clean_text)
print("Sesudah:", tokens)
```

Gambar 3. 12 Source code Tokenizing

Ilustrasi penulisan kode pada proses Tokenizing diperlihatkan dalam Gambar 3.12. Tahapan prapemrosesan ini diterapkan untuk memotong rangkaian teks menjadi potongan kata yang lebih kecil atau disebut token. Secara praktis, metode .split() pada Python dijalankan untuk mengurai teks dengan indikator berupa spasi. Mengingat teks sudah melalui fase cleaning dan case folding sebelumnya, seluruh komponen pengganggu seperti simbol, tanda baca, serta karakter non-alfabet telah dieliminasi. Alhasil, luaran yang dihasilkan berupa unit kata yang bersih dan siap digunakan pada tahapan pemrosesan lanjutan.

Pada Gambar 3.12 ditunjukkan implementasi kode pada tahap *Tokenizing* yang bertujuan untuk memecah teks menjadi unit kata yang lebih kecil (token). Proses *Tokenizing* dilakukan dengan memanfaatkan metode split() pada Python, yaitu dengan memisahkan teks berdasarkan spasi. Sebelum tahap *Tokenizing* dilakukan, teks telah melalui proses case folding dan cleaning, sehingga karakter non-alfabet, tanda baca, serta simbol yang tidak relevan telah dihapus. Dengan demikian, token yang dihasilkan berupa kata-kata yang lebih bersih dan siap digunakan pada tahap selanjutnya.

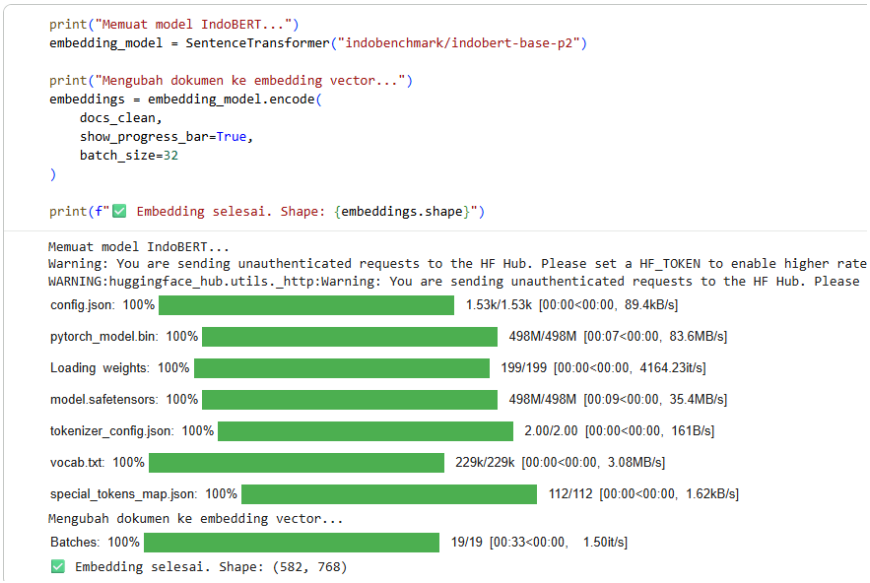
Selanjutnya, pada tahap representasi teks, proses tokenisasi utama dilakukan secara otomatis oleh tokenizer IndoBERT menggunakan mekanisme WordPiece. Dengan demikian, token hasil dari metode split() bukan merupakan representasi akhir, melainkan sebagai bentuk awal sebelum dilakukan proses embedding menggunakan IndoBERT. Berikut adalah tabel hasil preprocessing setelah seluruh tahapan pra-pemrosesan dilakukan.

Tabel 3. 7 Final Preprocessing

Text	Final_text
☕Warung kopi waris khas tulungagung... sangat menginspirasi... tempat nya sederhana... mantap👍👍👍	warung_kopi waris khas tulungagung sangat menginspirasi tempat nya sederhana mantap
Kopi Ijo" yang terkenal khususnya di wilayah Jawa Timur, lahir di tempat ini...	kopi_ijo terkenal khususnya wilayah jawa timur lahir tempat
Absen pagi menikmati kopi ijo diwaris yang sangat enak sekali!!!!!!👍👍👍	Absen pagi menikmati kopi_ijo diwaris sangat enak sekali

3.3 Implementasi Representasi Teks Dengan IndoBERT

Implementasi representasi teks dilakukan menggunakan model indobenchmark/indobert-base-p2 melalui pustaka SentenceTransformer pada Python. Setiap dokumen hasil preprocessing diproses untuk menghasilkan vektor embedding yang merepresentasikan makna dan konteks dari setiap ulasan. Vektor embedding tersebut kemudian digunakan sebagai masukan pada proses analisis topik menggunakan BERTopic.

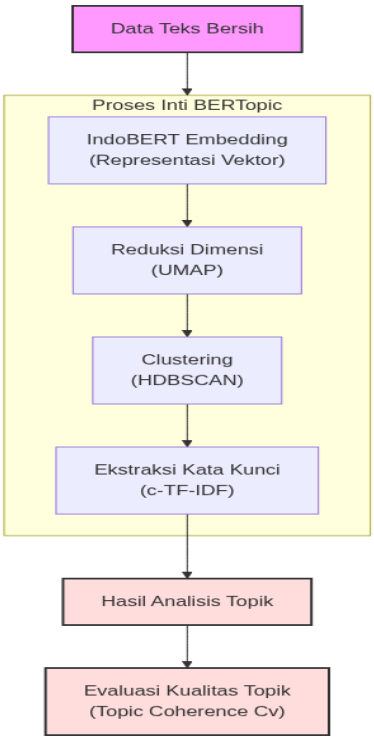


Gambar 3. 13 Source code BERTopic

Hasil implementasi menunjukkan bahwa sebanyak 582 dokumen berhasil direpresentasikan menjadi vektor embedding berdimensi 768. Jumlah dokumen tersebut sesuai dengan data hasil preprocessing yang digunakan dalam penelitian, sedangkan dimensi 768 merupakan ukuran representasi vektor yang dihasilkan oleh model `indobenchmark/indobert-base-p2`. Vektor embedding tersebut kemudian digunakan sebagai input pada proses analisis topik menggunakan BERTopic.

3.4 Implementasi Proses Analisis Topik

Analisis topik dalam penelitian ini dieksekusi dengan mengimplementasikan kerangka kerja BERTopic. Metode tersebut mengadopsi pendekatan *topic modelling* berbasis *unsupervised learning*, yang secara simultan mengintegrasikan serangkaian teknik mulai dari tahap *embedding*, reduksi dimensi, *clustering*, hingga pembobotan kosakata guna mengekstraksi klaster topik yang representatif.



Gambar 3. 14 Alur BERTopic

Berlandaskan pada kerangka metodologi yang telah dijabarkan pada Bab III, proses pemodelan topik dalam penelitian ini dilewati melalui beberapa tahapan sistematis. Alur pengerjaannya meliputi penerapan BERTopic, minimalisasi dimensi data lewat UMAP, pembentukan *cluster* dengan HDBSCAN, kalkulasi

bobot kata via c-TF-IDF, serta pengukuran kualitas topik menggunakan metrik *topic coherence*.

1. BERTopic

BERTopic merupakan sebuah pemodelan topik yang digunakan untuk mengidentifikasi topik-topik yang tersembunyi. Pada penelitian ini bERTopic digunakan dalam mengelompokan ulasan pengguna berdasarkan kemiripan makma. Proses diawali dengan mengubah teks menjadi representasi vektor menggunakan model indoBERT.

```
topic_model = BERTopic(
    embedding_model="indobenchmark/indobert-base-p2",
    vectorizer_model=vectorizer_model,
    min_topic_size=17,
    top_n_words=9,
    calculate_probabilities=True,
    verbose=True
)
```

Gambar 3. 15 Source ocode BERTopic

Gambar 3. 15 menunjukkan implementasi penggunaan metode BERTopic dalam penelitian ini. Parameter embedding_model menggunakan indoBERT untuk menghasilkan representasi teks berbasis embedding. Parameter vectorizer_model digunakan untuk mengatur p roses pembobotan kata, sedangkan min_topic_size menunjukkan jumlah minimal dokumen dalam satu topik.

2. Reduksi Dimensi dengan UMAP

Pasca-representasi dokumen diubah ke bentuk vektor embedding via IndoBERT, langkah berikutnya adalah mereduksi dimensi data menggunakan algoritma Uniform Manifold Approximation and Projection (UMAP). Prosedur ini

ditujukan untuk menyederhanakan data berdimensi tinggi ke dalam ruang dimensi yang lebih rendah tanpa mengorbankan esensi informasi penting di dalamnya. UMAP bekerja dengan memproyeksikan data sedemikian rupa sehingga dokumen-dokumen yang memiliki kedekatan konteks akan tetap berada di posisi yang saling merapat. Pola spasial ini sangat krusial dalam memperlancar proses clustering guna menghasilkan kelompok data yang lebih tegas dan terstruktur.

Pemanfaatan UMAP dalam studi ini difungsikan untuk memangkas kompleksitas data, sehingga kinerja algoritma clustering HDBSCAN dapat berjalan dengan lebih maksimal. Di samping itu, kelebihan UMAP terletak pada kemampuannya menjaga struktur lokal maupun global data demi menjamin relasi semantik antar-dokumen tidak bias. Kendati proses reduksi dimensi ini tidak dideklarasikan secara gamblang di dalam baris kode, mekanismenya telah terintegrasi secara otomatis di dalam pipeline BERTopic. Walau demikian, UMAP memegang andil besar dalam mendongkrak kualitas luaran pemodelan topik. Konfigurasi utama pada algoritma ini dikendalikan oleh parameter n_neighbors untuk mengatur kuantitas tetangga terdekat, serta min_dist yang meregulasi jarak paling minimum antar-titik pada ruang dimensi rendah.

3. Clustering dengan HDBSCAN

Pada penelitian ini, setelah dilakukan proses reduksi dimensi menggunakan UMAP, tahap selanjutnya adalah proses clustering menggunakan algoritma HDBSCAN. Tahap ini bertujuan untuk mengelompokkan dokumen berdasarkan tingkat kemiripan yang dimiliki dalam ruang embedding yang telah direduksi. HDBSCAN merupakan algoritma clustering berbasis kepadatan yang mampu

mengelompokkan data berdasarkan distribusi kepadatan tertentu. Algoritma ini bekerja dengan mengidentifikasi area dengan kepadatan tinggi sebagai cluster, sementara data yang berada di luar cluster akan dianggap sebagai noise.

Dalam konteks penelitian ini, algoritma HDBSCAN diterapkan untuk mengelompokkan dokumen ulasan berdasarkan kemiripan makna kontekstual yang telah direpresentasikan lewat vektor embedding dan direduksi parameternya via UMAP. Berkas dokumen yang memperlihatkan kedekatan makna akan dihimpun ke dalam satu klaster yang sama, sehingga masing-masing klaster dapat didefinisikan sebagai representasi topik tertentu. Salah satu keunggulan utama dari HDBSCAN terletak pada kapabilitasnya untuk menentukan kuantitas kelompok secara otomatis tanpa menuntut penetapan jumlah klaster di awal. Kelebihan ini selaras dengan karakteristik pendekatan unsupervised learning yang diusung dalam studi ini, di mana total topik keseluruhan belum diketahui secara apriori. Nilai tambah lainnya, HDBSCAN mampu mendeteksi dokumen yang tidak memiliki korelasi kuat dengan kelompok mana pun sebagai noise, sehingga akurasi hasil pengklasteran menjadi lebih kredibel. Adapun parameter operasional krusial pada HDBSCAN ialah min_cluster_size, yang meregulasi jumlah data paling minimal untuk membentuk sebuah klaster.

4. Pembobotan kata dengan c-TF-IDF

Begitu fase pengklasteran menggunakan algoritma HDBSCAN rampung, agenda komputasi dilanjutkan ke tahap pembobotan kata lewat metode class-based Term Frequency-Inverse Document Frequency (c-TF-IDF). Integrasi metode ini ditujukan untuk menyaring jajaran kata kunci yang paling representatif dari setiap

kelompok topik yang berhasil diidentifikasi. Sebagai bentuk modifikasi dari formula TF-IDF konvensional, c-TF-IDF beroperasi pada level kelas atau klaster, di mana seluruh dokumen yang bernaung dalam topik yang sama akan dilebur ke dalam satu kesatuan korpus dokumen. Proses kalkulasi bobotnya mengeksplorasi perpaduan nilai Term Frequency (TF) dengan Inverse Class Frequency (ICF). Lewat skema matematika tersebut, kosakata yang berulang kali muncul pada topik tertentu namun tergolong langka di topik-topik lainnya akan dianugerahi nilai bobot yang jauh lebih masif.

Pada penelitian ini, nilai A pada perhitungan ICF merepresentasikan rata-rata jumlah kata dalam setiap topik. Berdasarkan hasil perhitungan dari data penelitian, diperoleh nilai A = 9, yang menunjukkan bahwa setiap topik direpresentasikan oleh rata-rata 9 kata kunci. Nilai ini sejalan dengan parameter top_n_words yang digunakan dalam model BERTopic, sehingga jumlah kata kunci yang dihasilkan untuk setiap topik tetap konsisten dan representatif.

Tabel 3. 8 Contoh data topik

Topik A	Kopi enak mantap

Berdasarkan tabel 4.3 terdiri dari beberapa kata yaitu “kopi”, “enak”, dan “mantap”. Kata-kata tersebut digunakan sebagai contoh dalam proses perhitungan c-TF-IDF.

1. Menghitung Frekuensi Kata

Tabel 3. 9 Frekuensi kata

Kata	frekuensi
Kopi	1
Enak	2

mantap	1
Total	4

2. Menghitung kata TF

$$tf_{t,c} = \frac{ft,c}{\sum ft,c}$$

Tabel 3. 10 Perhitungan TF

Kata	**Perhitungan**	**TF**
Kopi	1/4	0.25
Enak	2/4	0.50
mantap	1/4	0.25

3. Menentukan Nilai global

- $A = 9$
- tft = Frekuensi kata pada seluruh topik

Tabel 3. 11 Nilai Global

Kata	***tft***
Kopi	8
Enak	7
mantap	2

4. Menghitung Inverse Class Frequency (ICF)

$$icf_t = \log\left(1 + \frac{A}{tft}\right)$$

Tabel 3. 12 Perhitungan ICF

kata	**perhitungan**	**ICF**
Kopi	log(1 + 9/8)	0.32
Enak	log(1 + 9/7)	0.36
mantap	log(1 + 9/2)	0.74

5. Menghitung c-TF-IDF

$$W_{t,c} = tf_{t,c} \times log\left(1 + \frac{A}{tf_t}\right)$$

Tabel 3. 13 Perhitungan c-TF-IDF

Kata	TF	ICF	c-TF-IDF
Kopi	0.25	0.32	0.08
Enak	0.50	0.36	0.18
mantap	0.25	0.74	0.18

Berdasarkan hasil perhitungan tersebut, kata “enak” dan “mantap” memiliki nilai c-TF-IDF tertinggi, sehingga kedua kata tersebut dianggap paling representatif dalam Topik A. Hal ini menunjukkan bahwa metode c-TF-IDF mampu mengidentifikasi kata yang tidak hanya sering muncul dalam suatu topik, tetapi juga memiliki tingkat keunikan yang lebih tinggi dibandingkan kata lainnya. Dengan demikian, kata-kata dengan nilai c-TF-IDF tertinggi dapat digunakan sebagai representasi utama dalam menginterpretasikan topik yang dihasilkan.

5. Topic Coherence

Setelah model BERTopic sukses memetakan dokumen ke dalam klaster-klaster tematik, agenda krusial berikutnya adalah melakukan pengujian terhadap kualitas topik bentukan tersebut. Di dalam ranah pemodelan topik (topic modeling), instrumen evaluasi yang paling sering diandalkan untuk kebutuhan ini ialah topic coherence. Parameter metrik ini diarsiteki secara khusus untuk mengestimasi derajat kedekatan serta konsistensi makna antar-kata yang mengonstruksi sebuah topik, Semakin koheren suatu topik, semakin mudah bagi manusia untuk memahami inti pembicaraan yang direpresentasikan oleh topik tersebut. Secara teknis, topic coherence bekerja dengan memanfaatkan korpus referensi eksternal atau internal untuk menghitung derajat kemunculan bersama (co-occurrence) antara kata-kata kunci dalam suatu topik. Beberapa varian metrik coherence yang umum digunakan antara lain C_V, C_Umass, dan C_NPMI, yang masing-masing memiliki pendekatan perhitungan yang berbeda. Dalam penelitian ini, metrik coherence dipilih karena kemampuannya dalam memberikan gambaran objektif mengenai kualitas topik tanpa memerlukan label atau anotasi manual dari manusia—sebuah keuntungan yang sangat berharga dalam konteks unsupervised learning.

Dalam pendekatan unsupervised learning, peneliti tidak memiliki data berlabel yang dapat digunakan untuk memverifikasi apakah hasil pemodelan sudah benar atau tidak. Berbeda dengan pendekatan klasifikasi di mana akurasi dapat dihitung dengan membandingkan prediksi terhadap label sebenarnya, topic modeling tidak memiliki "jawaban benar" yang pasti. Oleh karena itu, topic

coherence menjadi salah satu instrumen yang paling andal untuk menilai apakah topik yang dihasilkan oleh model benar-benar bermakna atau sekadar artefak statistik semata. Dalam penelitian ini, nilai topic coherence akan menjadi acuan utama dalam proses seleksi topik. Topik-topik yang memiliki koherensi tinggi akan dianggap sebagai topik yang valid dan representatif, sementara topik dengan koherensi rendah akan dieliminasi atau digabungkan dengan topik lain.

BAB IV

HASIL DAN PEMBAHASAN

4.1 Data Penelitian

Data penelitian ini bersumber dari ulasan Google Maps dari salah satu warung kopi terkenal di Tulungagung. Pengumpulan data dilakukan melalui proses scraping ulasan. Seluruh ulasan yang diperoleh akan melalui proses seleksi untuk memastikan relevansi data sebelum memasuki tahapan analisis topik menggunakan BERTopic. Data dalam penelitian ini tidak melalui proses pelabelan kategori secara manual, melainkan dikelompokkan secara otomatis berdasarkan kemiripan semantik melalui representasi embedding indoBERT. Hal ini dilakukan agar sistem dapat mengidentifikasi tema-tema utama yang muncul secara natural dari opini pelanggan tanpa adanya bias dari pelabelan awal. Distribusi jumlah ulasan yang berhasil dikumpulkan disajikan pada tabel 4.1 Berikut:

Table 4. 1 Data penelitian

Total ulasan mentah	1892
Setelah Ulasan yang dihapus (Tidak Relevan)	735
Setelah preprocessing	583

4.2 Hasil Pemodelan topik BERTopic

Hasil pemodelan topik menggunakan metode BERTopic ditampilkan pada Tabel 4.9. Berdasarkan hasil pemodelan, diperoleh 6 topik utama serta 1 kelompok noise (Topik -1). Setiap topik memuat informasi mengenai jumlah dokumen, kata kunci utama, serta contoh ulasan yang merepresentasikan karakteristik masing-masing topik.

Table 4. 2 Informasi Topik

No	Topic	Count	Name	Representation	Representative_Docs
0	-1	27	-1_kopi_rasa_legend_enak	kopi,rasa,legend,enak,kopi legend,mana,kenang masa,kenang,ijo	nyaman dar minum kopi rasa kopi sungguh enak,joss tenan kurang lengkap rasa tulungagung indak mampir kesini,awal beli bubuk kemas kopi ijo cari maps legend pernah kesini
1	0	231	0_tempat_nyaman_ngopi_nongkrong	tempat,nyaman,ngopi,nongkrong,teman,mantap,enak,asik,coco	tempat asik nyaman obrol sama teman,tempat nongkrong,tempat nyaman nongkrong sama tema
2	1	139	1_kopi_warung_warung kopi_tempat	kopi,warung,warung kopi,tempat,tulungagung,ijo,kopi ijo,khas,rasa	kopi hijau khas tulungagung saji cara sederhana mampu beri cita rasa nikmat harga sangat kunjung
3	2	58	2_kopi_ijo_kopi ijo_ijo kopi	kopi,ijo,kopi ijo,ijo kopi,kopi ijo kopi,mantap,ijo kopi ijo,kopi ijonya,ijonya	kopi ijo,kopi ijo,kopi ijo
4	3	57	3_tulungagung_kopi_warung kopi_tulungagung warung kopi	tulungagung,kopi,warung kopi,tulungagung warung kopi,tulungagung warung,warung,tulungagung kopi,khas,khas tulungagung	warung kopi legend tulungagung, warung kopi legend tulungagung, warung kopi legend tulungagung
5	4	38	4_kopi_mantap_kopi mantap_mantap kopi	kopi,mantap,kopi mantap,mantap kopi,kopi mantap kopi,enak kopi,kopi enak,enak,mantab	kopi mantap,kopi mantap,kopi mantap
6	5	32	5_coffee_good_the_of	coffee,good,the,of,place,best,for,coffe,the best	good place for good coffe,the

Berdasarkan hasil pemodelan topik pada Tabel 4.9, diperoleh enam topik utama serta satu kelompok noise (Topik -1). Topik -1 merupakan kumpulan dokumen yang tidak dapat dikelompokkan secara optimal ke dalam topik tertentu karena memiliki tingkat kemiripan yang rendah dengan dokumen lainnya, sehingga dikategorikan sebagai noise atau outlier oleh algoritma HDBSCAN. Secara umum, keenam topik yang terbentuk merepresentasikan berbagai aspek yang dibahas pengguna dalam ulasan warung kopi, meliputi kenyamanan tempat dan aktivitas nongkrong (Topik 0), identitas warung kopi (Topik 1), produk khas kopi ijo (Topik 2), karakteristik warung kopi sebagai bagian dari identitas daerah Tulungagung (Topik 3), kualitas dan cita rasa kopi (Topik 4), serta ulasan berbahasa Inggris yang mencerminkan keberagaman pengguna (Topik 5). Hasil tersebut menunjukkan bahwa ulasan pengguna tidak hanya berfokus pada produk kopi, tetapi juga mencakup pengalaman pengunjung, identitas lokal, serta suasana sosial yang terbentuk di warung kopi.

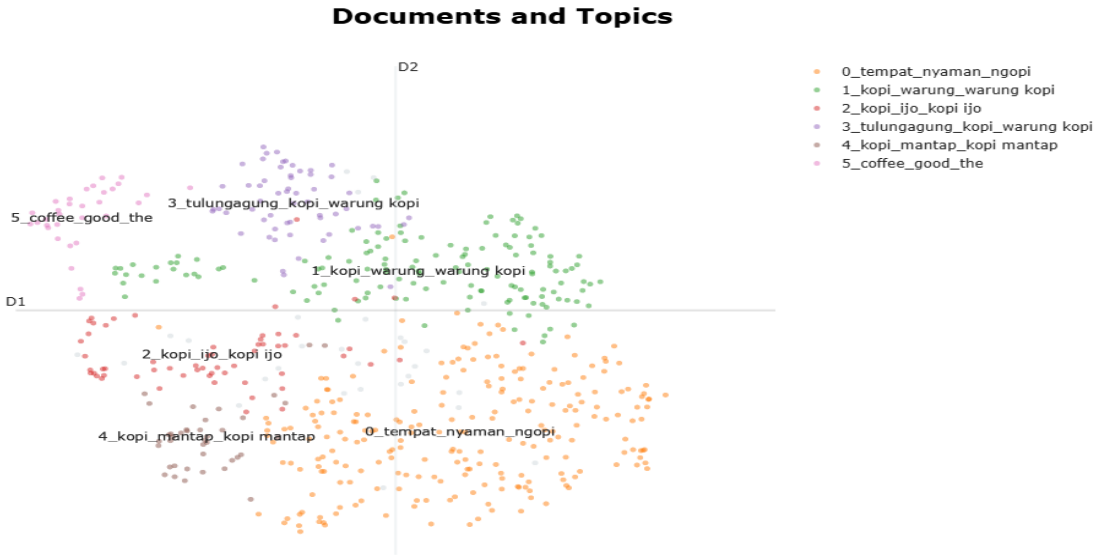
4.3 Visualisasi Topik

Visualisasi topik pada penelitian ini dilakukan untuk memberikan pemahaman yang lebih mendalam terhadap hasil pemodelan topik yang telah diperoleh. Melalui visualisasi, hubungan antar topik, distribusi topik, serta kata-kata yang merepresentasikan masing-masing topik dapat diamati secara lebih jelas dan terstruktur. Beberapa jenis visualisasi yang digunakan dalam penelitian ini meliputi intertopic distance map, topic word scores, wordcloud, serta distribusi topik. Setiap visualisasi memiliki peran yang berbeda dalam membantu interpretasi hasil, seperti menggambarkan kedekatan antar topik, menunjukkan kata-kata

dominan dalam topik, serta mengidentifikasi topik yang paling banyak dibahas oleh pengguna.

4.3.1 Intertopic Distance Map

Visualisasi Intertopic Distance Map digunakan untuk menggambarkan hubungan antar topik yang dihasilkan oleh metode BERTopic. Pada visualisasi ini, setiap lingkaran merepresentasikan satu topik, di mana ukuran lingkaran menunjukkan jumlah dokumen atau ulasan dalam topik tersebut. Sementara itu, jarak antar lingkaran menunjukkan tingkat kemiripan antar topik, di mana semakin dekat jaraknya maka semakin mirip konteks pembahasannya, dan semakin jauh menunjukkan perbedaan makna yang lebih signifikan.



Gambar 4. 1 Hasil Inttertopic Distance Map

Berdasarkan Gambar 4.1, terlihat bahwa topik-topik yang dihasilkan model BERTopic tersebar dalam ruang dua dimensi dengan posisi yang berbeda-beda.

Sebagian besar topik berada pada jarak yang relatif berdekatan, menunjukkan adanya keterkaitan konteks pembahasan yang didominasi oleh tema umum seperti kopi, warung kopi, dan pengalaman pelanggan. Di sisi lain, terdapat beberapa topik yang memiliki posisi lebih jauh dari kelompok utama, yang mengindikasikan adanya karakteristik pembahasan yang lebih spesifik dan berbeda dibandingkan topik lainnya, seperti pembahasan mengenai produk tertentu atau penggunaan kosakata yang khas dalam ulasan. Selain itu, ukuran lingkaran pada visualisasi menunjukkan proporsi jumlah dokumen pada setiap topik, di mana topik dengan lingkaran yang lebih besar memiliki jumlah ulasan yang lebih banyak dan menjadi topik yang lebih dominan dibandingkan topik lainnya. Kondisi ini menunjukkan bahwa terdapat tema-tema tertentu yang lebih sering dibahas oleh pengunjung warung kopi. Secara keseluruhan, hasil visualisasi Intertopic Distance Map menunjukkan bahwa model BERTopic mampu mengelompokkan data ulasan ke dalam beberapa topik yang memiliki karakteristik berbeda dengan tingkat keterkaitan semantik yang cukup baik, sehingga variasi tema yang terdapat pada data ulasan dapat direpresentasikan secara jelas.

4.3.2 Topic Word Scores

Visualisasi Topic Word Scores digunakan untuk menunjukkan kata-kata yang paling merepresentasikan setiap topik beserta bobot atau tingkat kepentingannya. Pada grafik ini, setiap topik ditampilkan dalam bentuk diagram batang yang berisi kata-kata utama dengan nilai skor tertentu. Semakin panjang batang yang ditampilkan, maka semakin tinggi kontribusi kata tersebut dalam membentuk topik.

Gambar 4. 2 Hasil Topic Word Scores

Berdasarkan hasil visualisasi Topic Word Scores, setiap topik memiliki kumpulan kata kunci dengan bobot yang berbeda-beda sehingga dapat menggambarkan tema utama dari masing-masing topik. Pada Topik 0, kata-kata seperti tempat, nyaman, ngopi, nongkrong, dan teman memiliki skor tertinggi, yang menunjukkan bahwa topik ini berkaitan dengan kenyamanan tempat dan aktivitas sosial pengunjung dalam menikmati kopi bersama teman atau keluarga. Hal ini mengindikasikan bahwa suasana warung kopi dapat menjadi salah satu aspek penilaian yang diberikan atau diperhatikan oleh pelanggan.

Pada Topik 1, kata-kata yang memiliki bobot tertinggi antara lain kopi, warung, warung kopi, tempat, dan Tulungagung. Kumpulan kata tersebut menggambarkan bahwa topik ini berkaitan dengan keberadaan warung kopi di Tulungagung sebagai tempat yang dikenal dan banyak dikunjungi oleh masyarakat. Topik ini merefleksikan persepsi umum pengunjung terhadap warung kopi sebagai bagian dari identitas daerah. Di sisi lain, Topik 2 didominasi oleh kata-kata seperti

kopi, ijo, dan kopi ijo. Dominasi kata-kata tersebut menunjukkan bahwa kopi ijo merupakan salah satu produk yang banyak mendapat perhatian dalam ulasan pengguna. Kemunculan istilah tersebut secara berulang mengindikasikan bahwa kopi ijo menjadi salah satu karakteristik khas yang melekat pada warung kopi di Tulungagung serta menjadi daya tarik bagi para pengunjung.

Topik 3 memiliki kata-kata dominan seperti Tulungagung, kopi, dan warung kopi, yang menggambarkan keterkaitan antara warung kopi dengan daerah asalnya serta popularitasnya di kalangan masyarakat. Pada Topik 4, kata-kata seperti kopi, mantap, dan kopi mantap menunjukkan adanya penilaian positif terhadap kualitas dan cita rasa kopi yang disajikan. Adapun Topik 5 didominasi oleh kata-kata berbahasa Inggris seperti coffee, good, dan place, yang menunjukkan bahwa sebagian pengguna memberikan ulasan menggunakan bahasa Inggris atau mencampurkan bahasa Indonesia dan bahasa Inggris dalam menyampaikan pendapatnya.

Secara keseluruhan, visualisasi Topic Word Scores menunjukkan bahwa metode BERTopic mampu mengidentifikasi kata-kata kunci yang representatif pada setiap topik. Perbedaan bobot kata pada masing-masing topik membantu dalam proses interpretasi tema yang terbentuk dan menunjukkan bahwa ulasan pengunjung tidak hanya membahas kualitas kopi, tetapi juga kenyamanan tempat, identitas lokal, produk khas, serta pengalaman yang diperoleh saat berkunjung ke warung kopi di Tulungagung.

4.3.3 WordCloud

Visualisasi WordCloud digunakan untuk memberikan gambaran umum mengenai kata-kata yang paling sering muncul dalam setiap topik hasil pemodelan. Pada visualisasi ini, ukuran kata mencerminkan tingkat frekuensi atau bobot kemunculan kata dalam suatu topik, di mana kata dengan ukuran lebih besar menunjukkan kata yang lebih dominan dibandingkan kata lainnya. WordCloud membantu dalam memahami pola kata secara visual dan intuitif, sehingga memudahkan dalam mengidentifikasi tema utama yang dibahas oleh pengguna. Dengan melihat kata-kata yang paling menonjol, peneliti dapat memperoleh gambaran cepat mengenai isi dan karakteristik masing-masing topik tanpa harus membaca seluruh data ulasan secara manual. Dalam penelitian ini, WordCloud ditampilkan untuk setiap topik yang dihasilkan oleh model BERTopic, sehingga dapat memperjelas hasil interpretasi topik yang telah dijelaskan sebelumnya. berikut merupakan visualisasi WordCloud untuk masing-masing topik yang telah diidentifikasi:

1. Topic 0

Gambar 4. 3 Topic 0

Berdasarkan visualisasi WordCloud pada Topik 0, terlihat bahwa kata tempat, ngopi, dan mantap merupakan kata yang paling sering dibahas pada hasil topik ini. Hal ini menunjukkan bahwa Topik 0 berkaitan erat dengan pengalaman pengguna dalam menikmati aktivitas minum kopi di suatu tempat. Selain itu, kemunculan kata seperti nyaman, teman, dan nongkrong mengindikasikan bahwa aspek kenyamanan dan fungsi sosial sebagai tempat berkumpul menjadi fokus utama dalam ulasan.

Kata kopinya dan enak juga muncul dalam visualisasi, yang menunjukkan adanya penilaian positif terhadap kualitas produk kopi yang disajikan. Namun, dibandingkan dengan kata lain, aspek suasana dan pengalaman berada di tempat tersebut lebih dominan dalam membentuk topik ini. Dengan demikian, Topik 0 dapat diinterpretasikan sebagai topik yang merepresentasikan pengalaman pelanggan terkait kenyamanan tempat dan aktivitas sosial seperti nongkrong bersama teman di warung kopi.

2. Topik 1

Gambar 4. 4 Topic 1

Pada Gambar visualisasi WordCloud pada Topik 1, terlihat bahwa kata kopi, tulungagung, dan warung merupakan kata yang paling dominan. Hal ini menunjukkan bahwa Topik 1 berkaitan dengan identitas warung kopi di wilayah Tulungagung. Kata-kata tersebut menggambarkan bahwa ulasan pengguna banyak membahas keberadaan warung kopi sebagai bagian dari karakteristik daerah.

Selain itu, munculnya kata seperti warung kopi, kota, dan khas mengindikasikan bahwa pengguna juga menyoroti ciri khas lokal serta posisi warung kopi sebagai bagian dari identitas kota. Kata kopi_ijo dan waris turut muncul, yang menunjukkan adanya pembahasan mengenai produk khas serta nama atau identitas warung kopi tertentu. Dengan demikian, Topik 1 dapat diinterpretasikan sebagai topik yang merepresentasikan identitas, lokasi, dan karakteristik khas warung kopi di Tulungagung.

3. Topik 2

Gambar 4. 5 Hasil Topik 2

Visualisasi WordCloud pada Topik 2, terlihat bahwa kata kopi_ijo dan kopi_ijonya merupakan kata yang paling dominan. Hal ini menunjukkan bahwa Topik 2 secara jelas berfokus pada pembahasan mengenai produk khas, yaitu kopi

ijo, yang menjadi ciri utama dari warung kopi tersebut. Selain itu, kemunculan kata mantap yang cukup dominan mengindikasikan adanya penilaian positif dari pengguna terhadap kualitas dan cita rasa kopi ijo. Kata spesialis juga muncul dalam visualisasi, yang menunjukkan bahwa warung kopi tersebut dikenal sebagai tempat yang memiliki keunggulan khusus dalam menyajikan kopi ijo. Dengan demikian, Topik 2 dapat diinterpretasikan sebagai topik yang merepresentasikan produk unggulan berupa kopi ijo serta persepsi positif pelanggan terhadap kualitas dan keunikan produk tersebut.

4. Topik 3

Gambar 4. 6 Hasil Topik 3

Dari hasil WordCloud pada Topik 3, terlihat bahwa kata kopi, legend, dan warung merupakan kata yang paling dominan. Hal ini menunjukkan bahwa Topik 3 berkaitan dengan reputasi warung kopi sebagai tempat yang sudah dikenal luas atau memiliki nilai historis sebagai warung kopi legendaris. Selain itu, kemunculan kata seperti warung kopi, kopi legend, dan kopi legendaris semakin memperkuat

bahwa topik ini membahas citra dan popularitas warung kopi di mata pelanggan. Penggunaan kata legend menunjukkan bahwa warung kopi tersebut dianggap memiliki keunikan, sejarah, atau kualitas yang telah diakui oleh banyak pengunjung.

5. Topik 4

Gambar 4. 7 Hasil Topik 4

Berdasarkan visualisasi WordCloud pada Topik 4, terlihat bahwa kata mantap dan kopi merupakan kata yang paling dominan dibandingkan kata lainnya. Dominasi kedua kata tersebut menunjukkan bahwa topik ini berkaitan dengan penilaian positif pengguna terhadap kualitas dan cita rasa kopi yang disajikan di warung kopi. Selain itu, muncul pula kata-kata seperti enak, kopi mantap, dan enak kopi yang semakin memperkuat bahwa ulasan pada topik ini berfokus pada kepuasan pengguna terhadap rasa kopi.

Ukuran kata yang lebih besar pada visualisasi menunjukkan frekuensi kemunculan dan tingkat kepentingan kata tersebut dalam topik. Semakin besar ukuran kata, maka semakin tinggi kontribusinya dalam membentuk topik tersebut. Dengan demikian, dapat disimpulkan bahwa Topik 4 merepresentasikan sentimen

positif pengguna terhadap kualitas kopi, terutama terkait rasa kopi yang dianggap enak dan mantap oleh pelanggan.

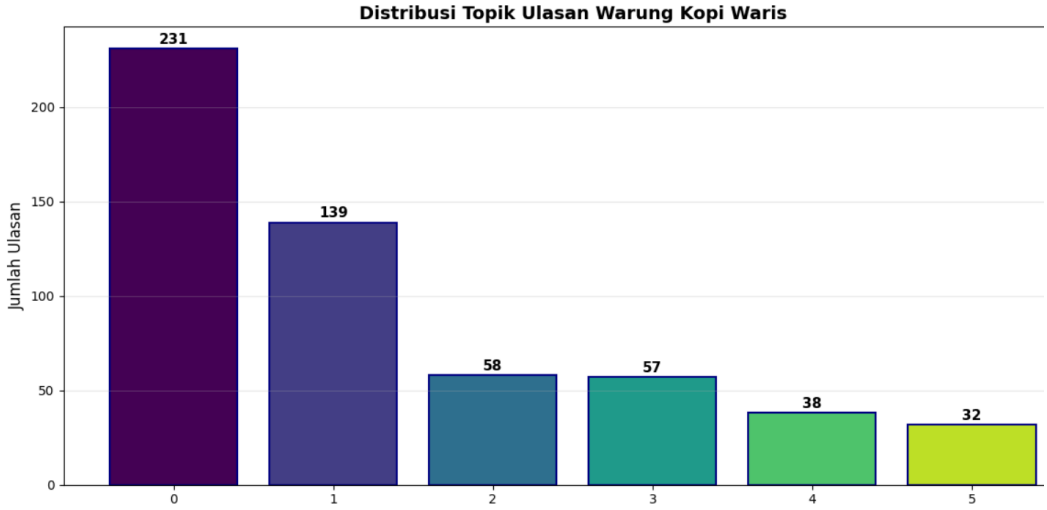
6. Topik 5

Gambar 4. 8 Hasil Topik 5

Berdasarkan visualisasi WordCloud pada Topik 4, terlihat bahwa kata coffee, good, dan best merupakan kata yang paling dominan. Hal ini menunjukkan bahwa Topik 4 berkaitan dengan penilaian positif terhadap kualitas kopi dan pengalaman pengguna, khususnya yang disampaikan dalam bahasa Inggris. Selain itu, kemunculan kata seperti place, the best, dan for mengindikasikan bahwa ulasan dalam topik ini tidak hanya membahas kualitas kopi, tetapi juga menilai tempat sebagai lokasi yang baik untuk menikmati kopi. Penggunaan bahasa Inggris pada kata-kata tersebut menunjukkan adanya variasi bahasa dalam ulasan pengguna, yang kemungkinan berasal dari pengguna non-Indonesia atau penggunaan code-mixing.

4.3.4 Distribusi Topik

Untuk mengetahui sebaran jumlah ulasan pada masing-masing topik yang telah dihasilkan, dilakukan visualisasi distribusi topik menggunakan diagram batang. Visualisasi ini bertujuan untuk memberikan gambaran mengenai topik mana yang paling dominan serta topik yang memiliki jumlah ulasan paling sedikit berdasarkan hasil pemodelan BERTopic.



Gambar 4. 9 Distribusi Topik

Berdasarkan hasil visualisasi distribusi topik pada Gambar 4.16, terlihat bahwa Topik 0 memiliki jumlah ulasan paling banyak dibandingkan dengan topik lainnya, sehingga dapat dikatakan sebagai topik yang paling dominan dalam data penelitian. Dominasi Topik 0 menunjukkan bahwa tema yang direpresentasikan oleh topik tersebut merupakan aspek yang paling sering dibahas oleh pengunjung warung kopi. Mengacu pada hasil interpretasi topik sebelumnya, Topik 0 berkaitan dengan kenyamanan tempat, aktivitas nongkrong, dan interaksi sosial, sehingga dapat diindikasikan bahwa pengunjung tidak hanya memberikan perhatian pada kualitas kopi, tetapi juga pada suasana dan pengalaman saat berkunjung.

Selanjutnya, Topik 1 memiliki jumlah ulasan yang cukup besar meskipun masih berada di bawah Topik 0. Topik ini menggambarkan identitas umum warung kopi di Tulungagung, sehingga banyaknya jumlah ulasan pada topik tersebut menunjukkan bahwa lokasi dan karakteristik warung kopi juga menjadi perhatian utama pengguna. Sementara itu, Topik 2 memiliki jumlah ulasan yang lebih sedikit, namun tetap cukup sering muncul dalam data. Topik ini berkaitan dengan pembahasan mengenai kopi ijo sebagai salah satu produk khas yang menjadi ciri tersendiri dari warung kopi di Tulungagung.

Di sisi lain, Topik 3 dan Topik 4 memiliki jumlah ulasan yang relatif lebih sedikit dibandingkan topik lainnya. Hal ini menunjukkan bahwa tema-tema tersebut hanya dibahas oleh sebagian pengguna dan cenderung bersifat lebih spesifik. Selain itu, Topik 5 juga memiliki proporsi yang lebih kecil, yang mengindikasikan bahwa penggunaan bahasa Inggris atau campuran bahasa dalam ulasan hanya dilakukan oleh sebagian kecil pengunjung.

Secara keseluruhan, distribusi jumlah ulasan pada setiap topik menunjukkan bahwa pembahasan pengguna tidak tersebar secara merata. Beberapa topik, terutama Topik 0 dan Topik 1, menjadi tema yang paling dominan, sedangkan topik lainnya muncul dalam jumlah yang lebih terbatas sesuai dengan karakteristik pembahasannya. Kondisi ini menunjukkan bahwa metode BERTopic mampu mengidentifikasi tema-tema utama yang paling banyak dibicarakan sekaligus membedakan topik-topik yang lebih spesifik dalam data ulasan pengunjung warung kopi di Tulungagung.

4.3.5 Topic Coherence

Untuk mengevaluasi kualitas topik yang dihasilkan oleh model BERTopic, digunakan metrik topic coherence dengan pendekatan C_v. Nilai coherence ini berfungsi untuk mengukur tingkat keterkaitan antar kata dalam satu topik, sehingga dapat menunjukkan apakah topik yang terbentuk memiliki makna yang jelas dan konsisten. Semakin tinggi nilai coherence, maka semakin baik kualitas topik yang dihasilkan oleh model.

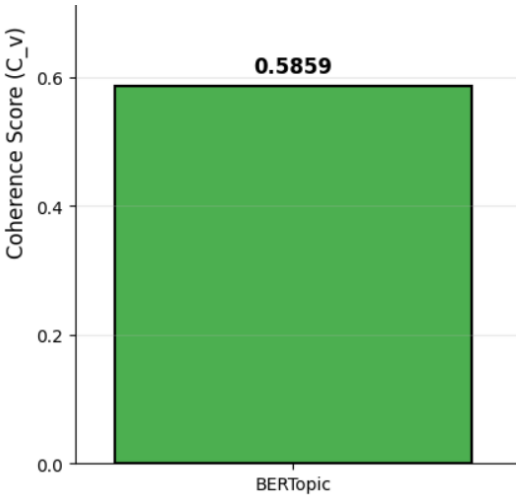
Berdasarkan Tabel 5.0, nilai coherence pada setiap topik menunjukkan variasi yang berbeda. Topik 3 memiliki nilai coherence tertinggi sebesar 0,8205 dengan kata kunci utama "tulungagung", "kopi", dan "warung kopi". Nilai tersebut menunjukkan bahwa kata-kata dalam topik ini memiliki keterkaitan semantik yang sangat kuat sehingga topik dapat diinterpretasikan dengan jelas. Sebaliknya, Topik 4 memiliki nilai coherence terendah sebesar 0,3844 dengan kata kunci utama "kopi", "mantap", dan "kopi mantap". Meskipun memiliki nilai lebih rendah dibandingkan topik lainnya, topik tersebut masih mampu menggambarkan tema tertentu yang muncul dalam ulasan pengguna.

Selain itu, Topik 1 dan Topik 2 juga memperoleh nilai coherence yang cukup tinggi, masing-masing sebesar 0,6576 dan 0,6483. Hal ini menunjukkan bahwa topik-topik tersebut memiliki hubungan antar kata yang cukup baik dan dapat merepresentasikan tema yang konsisten. Sementara itu, Topik 0 dan Topik 5 memperoleh nilai coherence sebesar 0,4020 dan 0,6024, yang menunjukkan tingkat keterkaitan kata yang masih dapat diinterpretasikan dengan baik

Table 4. 3 Nilai Coherence BERTopic

Topik	Kata Kunci Utama	Coherence Score
Topik 0	tempat, nyaman, ngopi, nongkrong, teman	0.4020
Topik 1	kopi, warung, warung kopi, tempat, tulungagung	0.6576
Topik 2	kopi, ijo, kopi ijo, ijo kopi, kopi ijo kopi	0.6483
Topik 3	tulungagung, kopi, warung kopi, tulungagung warung kopi	0.8205
Topik 4	kopi, mantap, kopi mantap, mantap kopi	0.3844
Topik 5	coffee, good, the, of, place	0.6024

Pada Gambar 4.10 juga memperlihatkan nilai topic coherence *(C_v)* yang dihasilkan oleh model BERTopic adalah sebesar 0,5859. Nilai tersebut menunjukkan bahwa kata-kata dalam setiap topik memiliki tingkat keterkaitan yang cukup baik dan tidak terbentuk secara acak. Secara umum, nilai coherence yang mendekati 1 menunjukkan kualitas topik yang semakin baik, sedangkan nilai di atas 0,5 umumnya telah dianggap cukup untuk menghasilkan topik yang dapat diinterpretasikan dengan baik. Oleh karena itu, nilai coherence sebesar 0,5859 menunjukkan bahwa model BERTopic mampu menghasilkan topik yang koheren dan representatif terhadap data ulasan yang dianalisis.



Gambar 4. 10 Topic Coherence BERTopic

Hasil ini menunjukkan bahwa metode BERTopic yang dikombinasikan dengan embedding IndoBERT mampu menghasilkan topik yang relevan, memiliki hubungan semantik yang kuat antar kata, serta mudah diinterpretasikan dalam proses analisis topik.

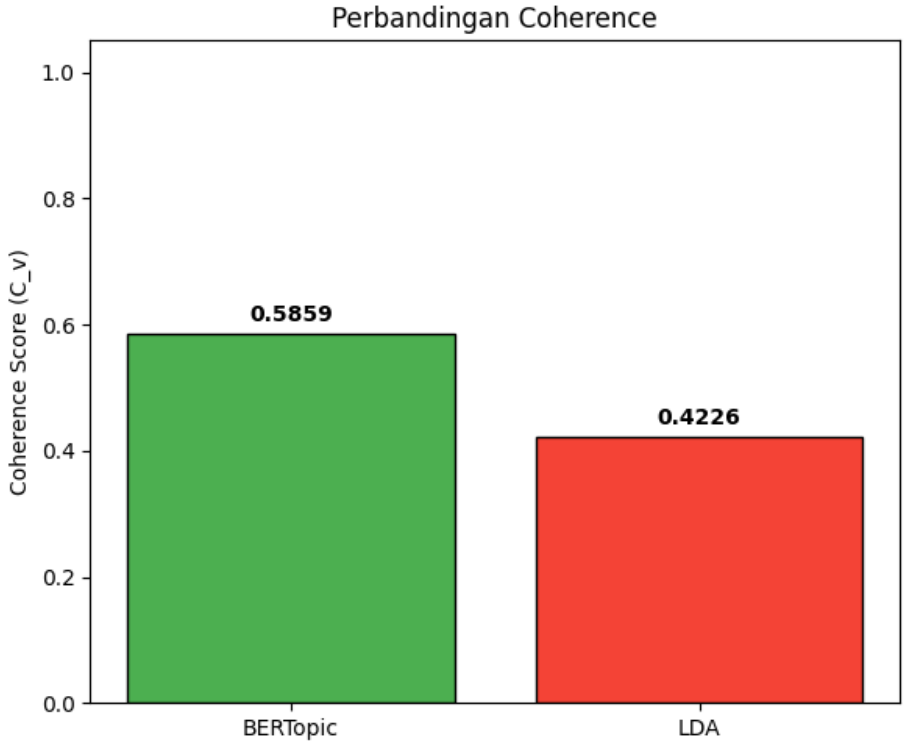
4.3.6 Perbandingan BERTopic dan LDA

Untuk mengevaluasi kualitas topik yang dihasilkan oleh model BERTopic dan LDA, digunakan metrik topic coherence (C_v). Metrik ini digunakan untuk mengukur tingkat keterkaitan semantik antar kata dalam suatu topik, sehingga dapat menunjukkan seberapa baik topik yang dihasilkan dapat dipahami secara makna (Röder & Hinneburg, n.d.).

Table 4. 4 Perbandingan Coherence Global Model BERTopic dan LDA

Metode	Coherence Score (C_v)
BERTopic	0,5859
LDA	0.4226
Selisih	0.1633

Pada tabel 4.4 menyajikan tabel ringkasan perbandingan coherence global antara model BERTopic dan LDA. Berdasarkan tabel tersebut, terlihat bahwa BERTopic memperoleh nilai coherence sebesar 0,5859, sedangkan LDA hanya mencapai 0,4226. Selisih sebesar +0,1633 atau +38,6% menunjukkan bahwa BERTopic secara keseluruhan menghasilkan topik yang lebih koheren dibandingkan LDA.

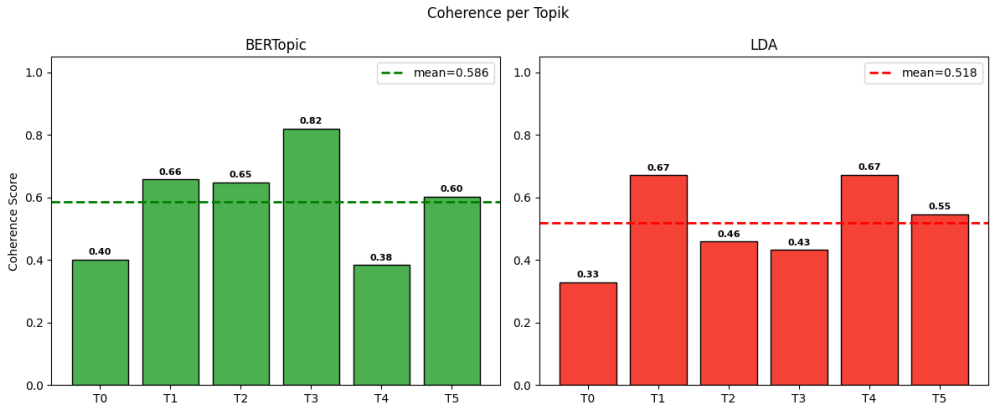


Gambar 4. 11 Perbandingan Score Coherence

Berdasarkan Gambar 4.11, terlihat bahwa model BERTopic memperoleh nilai topic coherence (C_v) sebesar 0,5859, sedangkan model LDA memperoleh nilai sebesar 0,4226. Perbedaan nilai tersebut menunjukkan bahwa BERTopic menghasilkan topik dengan tingkat keterkaitan semantik antar kata yang lebih baik dibandingkan LDA pada data ulasan warung kopi yang dianalisis.

Nilai coherence yang lebih tinggi mengindikasikan bahwa kata-kata dalam setiap topik memiliki hubungan yang lebih konsisten sehingga topik yang dihasilkan lebih mudah diinterpretasikan. Hasil ini menunjukkan bahwa pendekatan BERTopic yang memanfaatkan representasi embedding mampu membentuk topik yang lebih representatif terhadap isi ulasan dibandingkan pendekatan LDA.

Meskipun demikian, nilai coherence global hanya memberikan gambaran umum mengenai kualitas model secara keseluruhan. Oleh karena itu, untuk memperoleh analisis yang lebih mendalam, dilakukan evaluasi coherence pada masing-masing topik yang dihasilkan oleh kedua model. Analisis tersebut bertujuan untuk mengetahui kontribusi setiap topik terhadap kualitas model secara keseluruhan serta melihat topik mana yang memiliki tingkat keterkaitan semantik paling baik.



Gambar 4. 12 Perbandingan Score Coherence Per Topik

Gambar 4.12 menyajikan perbandingan nilai topic coherence pada setiap topik yang dihasilkan oleh model BERTopic dan LDA. Berdasarkan visualisasi tersebut, terlihat bahwa BERTopic memperoleh nilai coherence yang lebih tinggi pada lima dari enam topik yang dihasilkan, sedangkan LDA hanya memperoleh nilai yang sedikit lebih tinggi pada Topik 1. Keunggulan terbesar BERTopic terlihat pada Topik 3 dengan nilai coherence sebesar 0,8205, sedangkan LDA memperoleh nilai sebesar 0,3020. Hasil tersebut menunjukkan bahwa kata-kata yang membentuk Topik 3 pada BERTopic memiliki keterkaitan semantik yang lebih kuat sehingga topik yang dihasilkan lebih mudah diinterpretasikan. Sementara itu, pada Topik 1

perbedaan nilai coherence antara kedua model relatif kecil, sehingga kualitas topik yang dihasilkan dapat dikatakan hampir sebanding. Secara umum, visualisasi ini menunjukkan bahwa BERTopic mampu menghasilkan topik dengan tingkat keterkaitan antar kata yang lebih baik pada sebagian besar topik yang terbentuk. Untuk memperoleh gambaran yang lebih rinci mengenai perbedaan nilai coherence pada setiap topik, hasil tersebut selanjutnya disajikan pada Tabel 4.5.

Table 4. 5 Perbandingan Coherence pertopik

Topik	BERTopic	LDA	Selisih
Topik 0	0,4020	0,3295	+0,0725
Topik 1	0,6576	0,6709	-0,0133
Topik 2	0,6483	0,4589	+0,1894
Topik 3	0,8205	0,3020	+0,5185
Topik 4	0,3844	0,3020	+0,0824
Topik 5	0,6024	0,3020	+0,3004

Untuk memperkuat hasil evaluasi kuantitatif, Tabel 4.5 menyajikan perbandingan nilai topic coherence pada setiap topik yang dihasilkan oleh model BERTopic dan LDA. Berdasarkan tabel tersebut, BERTopic memperoleh nilai coherence yang lebih tinggi pada lima dari enam topik yang dihasilkan. Keunggulan terbesar terdapat pada Topik 3, di mana BERTopic memperoleh nilai coherence sebesar 0,8205, sedangkan LDA memperoleh nilai sebesar 0,3020 dengan selisih sebesar 0,5185. Hasil tersebut menunjukkan bahwa kata-kata penyusun Topik 3 pada BERTopic memiliki keterkaitan semantik yang lebih kuat sehingga topik yang dihasilkan lebih mudah diinterpretasikan.

Selain itu, BERTopic juga menunjukkan keunggulan yang cukup besar pada Topik 2 dan Topik 5, sedangkan pada Topik 0 dan Topik 4 perbedaannya relatif lebih kecil. Di sisi lain, LDA hanya memperoleh nilai coherence yang sedikit lebih tinggi pada Topik 1 dengan selisih sebesar 0,0133, sehingga kualitas topik yang dihasilkan kedua model pada topik tersebut dapat dikatakan hampir sebanding. Secara keseluruhan, hasil perbandingan coherence per topik menunjukkan bahwa BERTopic mampu menghasilkan topik dengan tingkat keterkaitan antar kata yang lebih baik pada sebagian besar topik yang terbentuk. Hasil ini sejalan dengan pembahasan topik dan visualisasi yang telah dijelaskan sebelumnya, di mana BERTopic mampu menghasilkan tema-tema yang lebih jelas dan mudah diinterpretasikan pada data ulasan warung kopi di Tulungagung.

4.4 Pembahasan

Berdasarkan hasil evaluasi menggunakan metrik topic coherence (C_v), model BERTopic berbasis IndoBERT memperoleh nilai coherence sebesar 0,5859, sedangkan model LDA memperoleh nilai 0,4226. Perbedaan nilai tersebut menunjukkan bahwa BERTopic mampu membentuk topik dengan keterkaitan semantik antar kata yang lebih baik dibandingkan LDA. Semakin tinggi nilai coherence, semakin baik pula hubungan antar kata dalam suatu topik sehingga tema yang dihasilkan menjadi lebih jelas dan mudah dipahami.

Temuan tersebut diperkuat oleh hasil evaluasi coherence pada masing-masing topik. BERTopic menghasilkan nilai coherence yang lebih tinggi pada lima dari enam topik yang terbentuk. Keunggulan paling besar terlihat pada Topik 3, di mana BERTopic mencapai nilai 0,8205, sedangkan LDA memperoleh nilai 0,3020. Hasil

ini mengindikasikan bahwa BERTopic mampu mengelompokkan kata-kata dengan hubungan semantik yang lebih kuat sehingga menghasilkan topik yang lebih representatif. Di sisi lain, LDA hanya sedikit lebih baik pada Topik 1 dengan selisih yang relatif kecil, sehingga performa kedua metode pada topik tersebut dapat dianggap hampir setara.

Perbedaan performa tersebut dipengaruhi oleh pendekatan yang digunakan oleh masing-masing metode. BERTopic memanfaatkan representasi embedding berbasis IndoBERT yang mampu menangkap hubungan kontekstual antar kata dalam suatu kalimat, sedangkan LDA membentuk topik berdasarkan distribusi probabilitas kemunculan kata dalam dokumen. Akibatnya, BERTopic mampu menghasilkan topik yang lebih koheren pada data ulasan warung kopi yang memiliki variasi bahasa dan beragam konteks pembahasan. Selain itu, hasil evaluasi coherence ini juga sejalan dengan hasil visualisasi Intertopic Distance Map, Topic Word Scores, WordCloud, dan Distribusi Topik yang telah dibahas sebelumnya. Topik-topik yang dihasilkan BERTopic mampu menggambarkan tema-tema utama dalam data ulasan, seperti suasana warung kopi, kopi ijo khas Tulungagung, identitas warung kopi, kualitas produk, serta aktivitas sosial pengunjung. Hal ini menunjukkan bahwa BERTopic tidak hanya menghasilkan nilai evaluasi yang baik secara kuantitatif, tetapi juga menghasilkan topik yang mudah diinterpretasikan secara kualitatif. Secara keseluruhan, hasil penelitian menunjukkan bahwa metode BERTopic berbasis IndoBERT memiliki performa yang lebih baik dibandingkan LDA dalam proses ekstraksi topik ulasan warung kopi di Tulungagung. Keunggulan tersebut terlihat dari nilai coherence global yang lebih tinggi serta

kualitas topik yang lebih baik pada sebagian besar topik yang dihasilkan. Dengan demikian, BERTopic berbasis IndoBERT dapat menjadi pendekatan yang efektif dalam proses ekstraksi topik pada data ulasan berbahasa Indonesia.

Dalam perspektif Islam, manusia dianjurkan untuk menjalin hubungan sosial yang baik dengan sesama. Allah SWT berfirman dalam Al-Qur'an surat Al-Hujurat ayat 13:

يَـٰٓأَيُّهَا ٱلنَّاسُ إِنَّا خَلَقْنَـٰكُم مِّن ذَكَرٍ وَأُنثَىٰ وَجَعَلْنَـٰكُمْ شُعُوبًا وَقَبَآئِلَ لِتَعَارَفُوٓاإِنَّ أَكْرَمَكُمْ عِندَ ٱللَّهِ أَتْقَىٰكُمْ إِنَّ ٱللَّهَ عَلِيمٌ خَبِيرٌ

" Wahai manusia, sesungguhnya Kami telah menciptakan kamu dari seorang laki-laki dan perempuan. Kemudian, Kami menjadikan kamu berbangsa-bangsa dan bersuku-suku agar kamu saling mengenal. Sesungguhnya yang paling mulia di antara kamu di sisi Allah adalah orang yang paling bertakwa. Sesungguhnya Allah Maha Mengetahui lagi Mahateliti."(Qs. Al-Hujarat:13)

Ayat tersebut menjelaskan bahwa manusia diciptakan untuk saling mengenal, berinteraksi, dan membangun hubungan sosial yang harmonis. Hal ini sejalan dengan hasil penelitian yang menunjukkan adanya topik mengenai aktivitas "nongkrong", "teman", dan interaksi sosial di warung kopi. Aktivitas tersebut mencerminkan fungsi warung kopi sebagai tempat berkumpul dan mempererat hubungan sosial masyarakat. Selain itu, Islam juga menganjurkan umatnya untuk menjaga dan mempererat hubungan sosial melalui silaturahmi. Rasulullah SAW bersabda:

"Siapa yang ingin dilapangkan rezekinya dan dipanjangkan umurnya, maka hendaklah ia menyambung silaturahmi." (HR. Bukhari No. 5986 dan Muslim No. 2557)

Hadis tersebut menunjukkan pentingnya menjaga hubungan sosial antarmanusia. Dalam konteks penelitian ini, warung kopi dapat menjadi media

untuk mempererat silaturahmi, berdiskusi, dan membangun komunikasi yang baik antarindividu selama dilakukan dalam kegiatan yang positif dan tidak bertentangan dengan syariat Islam.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa metode BERTopic berbasis IndoBERT mampu melakukan analisis topik pada data ulasan warung kopi di Tulungagung dengan hasil yang cukup baik. Dapat disimpulkan sebagai berikut:

1. Model IndoBERT berhasil diterapkan sebagai representasi embedding teks ulasan pengunjung warung kopi Waris di Tulungagung. Proses embedding menghasilkan representasi vektor kontekstual yang mampu menangkap makna semantik antar kata dalam bahasa Indonesia, termasuk bahasa informal dan variasi bahasa yang umum ditemukan pada data ulasan Google Maps.
2. Metode BERTopic berbasis IndoBERT berhasil mengidentifikasi enam topik utama dari ulasan pengunjung warung kopi Waris di Tulungagung, yang berkaitan dengan kenyamanan tempat dan aktivitas nongkrong, identitas warung kopi di Tulungagung, produk khas kopi ijo, reputasi warung kopi yang legendaris, serta ulasan berbahasa Inggris. Topik-topik tersebut menggambarkan bahwa warung kopi tidak hanya berperan sebagai tempat konsumsi, tetapi juga sebagai ruang interaksi sosial masyarakat.
3. berdasarkan evaluasi menggunakan metrik topic coherence (*Cv*), metode BERTopic memperoleh nilai sebesar 0,5859, lebih tinggi dibandingkan metode LDA yang memperoleh nilai sebesar 0,4226. Hasil ini menunjukkan

bahwa BERTopic mampu menghasilkan topik yang lebih koheren dan lebih mudah diinterpretasikan dibandingkan LDA pada data ulasan Google Maps berbahasa Indonesia. Dengan demikian, BERTopic berbasis IndoBERT dapat menjadi alternatif yang efektif untuk analisis topik pada data ulasan pengguna berbahasa Indonesia.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, masih terdapat beberapa keterbatasan yang dapat menjadi bahan pengembangan pada penelitian selanjutnya. Berikut beberapa saran yang dapat dipertimbangkan untuk meningkatkan kualitas penelitian di bidang analisis topik berbasis NLP:

1. Data yang digunakan dalam penelitian ini hanya berasal dari ulasan Google Maps pada Warung Kopi Waris di Tulungagung. Penggunaan data dari beberapa warung kopi atau platform ulasan lainnya berpotensi menghasilkan topik yang lebih beragam dan memberikan gambaran yang lebih luas mengenai preferensi pelanggan.
2. Perbandingan metode dalam penelitian ini masih terbatas pada BERTopic dan LDA. Penggunaan metode topic modeling lain seperti Top2Vec atau Neural Topic Modeling dapat dilakukan untuk mengetahui metode yang paling sesuai dalam menganalisis ulasan berbahasa Indonesia.
3. Beberapa ulasan masih mengandung bahasa daerah, singkatan, kata tidak baku, serta campuran bahasa Indonesia dan Inggris. Penanganan yang lebih spesifik terhadap karakteristik bahasa tersebut dapat dilakukan untuk meningkatkan kualitas representasi teks dan hasil pemodelan topik.

4. Evaluasi dalam penelitian ini menggunakan metrik topic coherence (Cv). Penggunaan metrik evaluasi tambahan dapat dipertimbangkan agar kualitas topik yang dihasilkan dapat dianalisis dari berbagai sudut pandang.

DAFTAR PUSTAKA

Alsayat, A., & Elmitwally, A. (2020). A comprehensive study on text classification techniques. *IEEE Access*. https://doi.org/10.1109/ACCESS.2020.2992423

Anugrah, A. F., & Agatha, R. D. (2025). Utilizing IndoBERT and BERTopic to Explore Public Opinion on BPS Instagram Posts. *Journal (Unspecified)*, *9*(5), 2836–2844.

Apriyadi, C., & Styawati, S. (2025). Sentiment Analysis of Cyber Attacks in Bank Syariah Indonesia Using SVM and Indobert Method. *Jurnal Teknik Informatika (Jutif)*, *6*(2), 819–838. https://doi.org/10.52436/1.jutif.2025.6.2.2636

Arifianto, E. Y., Digital, K. F., Studi, P., Program, I., Industri, F. T., & Indonesia, U. I. (2020). *Analisis Topik Data Tindak Kriminal pada Media Sosial Twitter Menggunakan Metode LDA (Latent Dirichlet Allocation)*.

Chamid, A. A., Nindyasari, R., Azizah, N., & Hariyadi, A. (2025). Analysis of Public Opinion on The Governor Candidate Debate Using LDA and IndoBERT. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, *4*(3). https://doi.org/10.22219/kinetik.v10i3.2221

Costa, F. De, & Aziz, N. A. (2021). *The Effects of User Generated Content and Firm Generated Content on Millennials' Purchase Intention of Shariah-Compliant Stocks*. *62*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2020). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. *Mlm*.

Dwiyono, A., Abdiansah, A., & Fachrurrozi, M. (2024). Analisis Perbandingan Klasifikasi Intent Chatbot Menggunakan Deep Learning BERT, RoBERTa, dan IndoBERT. *Journal of Information System Research (JOSH)*, *6*(1), 595–606. https://doi.org/10.47065/josh.v6i1.6051

Egger, R., & others. (2022). Topic modeling with BERT embeddings: A case study of online reviews. *Journal of Business Research*, *144*, 117–130. https://doi.org/10.1016/j.jbusres.2022.01.019

Filieri, R., & others. (2021). The influence of electronic word of mouth on consumer decision-making. *Annals of Tourism Research*, *89*, 103252. https://doi.org/10.1016/j.annals.2021.103252

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv Preprint ArXiv:2203.05794*.

Jakhotiya, A., Jain, H., Jain, B., & Chaniyara, M. C. (2022). *Text Pre-Processing Techniques in Natural Language Processing : A Review*. 878–880.

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoBERT: A Pretrained Language Model for Indonesian. *Proceedings of the 28th International Conference on Computational Linguistics*, 757–770. https://doi.org/10.18653/v1/2020.coling-main.66

Li, J. et al. (2020). Text mining and natural language processing in social media analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. https://doi.org/10.1002/widm.1348

Nidhal, M., Indrayadi, S., Budiman, L., & Sutrisna, A. (n.d.). *Menelusuri Ulasan Online di Indonesia : Dampak pada Platform UGC dan Konsumen*. *62*.

Rachman, A., Setiawati, L., & Rafdinal, W. (2021). *Pengaruh Informasi UGC dari Media Sosial pada Minat Berwisata di Destinasi Wisata Jawa Barat*. 4–5.

Ramadhan, C., Atina, V., & Permatasari, H. (2025). Analisis Perbandingan Model CNN dan IndoBERT Dalam Sentimen Berita Politik Indonesia. *Prosiding Seminar Nasional Teknologi Informasi Dan Bisnis*, 110–118. https://doi.org/10.47701/v1r9ka69

Rangkuti, R. N., Mardhiyah, A., Siregar, O. M., & Harahap, K. (2025). *Pengaruh User Generated Content dan Brand Love terhadap Keputusan Pembelian (Studi pada Konsumen Produk Ms Glow di Kota Medan)*. *4*(1), 401–407.

Röder, M., & Hinneburg, A. (n.d.). *Exploring the Space of Topic Coherence Measures*. 399–408.

Rubyanti, D. (2020). *PERAN USER GENERATED CONTENT (UGC) INSTAGRAM PADA INDUSTRI MAKANAN*. *04*(01), 1–24.

Sihombing, W. M., & Widiyaningtyas, T. (2025). Sentiment Analysis and Topic Modelling Using IndoBERTweet and BERTopic for Public Health Issues. *Indonesian Journal of Innovation Studies*, *26*(4). https://doi.org/10.21070/ijins.v26i4.1833

Wayan, N., Suaryani, N., Putri, T., & Suhendro, J. M. (2025). *Analisis User Generated Content untuk Mengidentifikasi Topik Pembicaraan Menggunakan Topic Modelling*. *4*(4), 6690–6697.

Wilie, B. et al. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian NLP. *ArXiv Preprint ArXiv:2009.05387*.

Yan, H., & Zhang, Y. (2023). An optimized BERTopic framework based on cluster silhouette for improving topic coherence. *ACM International Conference Proceeding Series*, *April*, 559–566. https://doi.org/10.1145/3652628.3652721