

PEMODELAN TOPIK PADA TWEET BAHASA INDONESIA MENGUNAKAN BERTOPIC

Diajukan Sebagai Syarat Untuk Menyelesaikan
Pendidikan Program Strata-1 Pada
Jurusan Teknik Informatika



Oleh :

Fahmi Guntara Diyasa

NIM : 09021181823002

**Jurusan Teknik Informatika
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA
2023**

LEMBAR PENGESAHAN SKRIPSI

PEMODELAN TOPIK PADA TWEET BAHASA INDONESIA
MENGUNAKAN BERTOPIC

Oleh:

Fahmi Guntara Diyasa
NIM: 09021181823002

Indralaya, Agustus 2023

Mengetahui,
Ketua Jurusan Teknik Informatika



Alvi Syahrini Utami, M.Kom
NIP. 197812222006042003

Pembimbing,

Dr. Abdiansah, S.Kom., M.Cs.
NIP. 198410012009121005

TANDA LULUS UJIAN KOMPREHENSIF

Pada hari senin tanggal 31 juli 2023 telah dilaksanakan ujian komprehensif skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Fahmi Guntara Diyasa

NIM : 09021181823002

Judul : Pemodelan Topik Pada Tweet Bahasa Indonesia Menggunakan BERTopic

Dan dinyatakan LULUS

1. Ketua Penguji

Yunita, M.Cs.

NIP. 198306062015042002

2. Penguji

Novi Yusliani, M.T.

NIP. 198211082012122001

3. Pembimbing

Dr. Abdiansah, S.Kom., M.Cs.

NIP. 198410012009121005

Mengetahui

Ketua Jurusan Teknik Informatika



Alvi Syahrini Utami, M.Kom
NIP. 197812222006042003

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama : Fahmi Guntara Diyasa

NIM : 09021181823002

Program Studi : Teknik Informatika

Judul Skripsi : Pemodelan Topik Pada Tweet Bahasa Indonesia Menggunakan
BERTopic

Hasil pengecekan software iThenticate/Turnitin: 14%

Menyatakan bahwa laporan proyek saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan proyek ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan dari pihak manapun.



Indralaya, 14 Agustus 2023



Fahmi Guntara Diyasa
NIM. 09021181823002

MOTTO DAN PERSEMBAHAN

Motto:

Keep it simple

Kupersembahkan Karya Tulis ini kepada:

- Allah SWT
- Kedua orang tua, saudara dan teman saya
- Fakultas Ilmu Komputer
- Universitas Sriwijaya

ABSTRACT

The increasing use of information and communication technology in recent years has had a significant impact on the trend of communicating and expressing aspirations through social media, especially the Twitter platform. Every tweet on Twitter carries information about a particular topic that can be identified through the Topic Modeling method. Topic Modeling is a tool used to uncover hidden topics in a group of documents. This research aims to perform topic modeling on Indonesian tweets using BERTopic. The Topic Modeling process using BERTopic includes steps such as document embedding, dimension reduction with UMAP, document clustering using HDBSCAN, and representing topics using c-TF-IDF. The dataset used consists of 10,000 Indonesian tweets taken from the Twitter account @detikcom. From the 10,000 tweets, 119 main topics and 1 outlier topic were found. Topic Modeling Evaluation is done using coherence score cv, with the average coherence score cv of 0.685, the highest coherence score cv of 0.995, and the lowest coherence score cv of 0.119..

Keywords: *Topic Modeling, BERTopic, Cohrence Score cv, Tweets*

ABSTRAK

Peningkatan penggunaan teknologi informasi dan komunikasi dalam beberapa tahun terakhir telah membawa dampak signifikan terhadap tren berkomunikasi dan penyampaian aspirasi melalui media sosial, khususnya platform Twitter. Setiap *tweet* di Twitter membawa informasi mengenai topik tertentu yang dapat diidentifikasi melalui metode Pemodelan Topik. Pemodelan Topik merupakan alat yang digunakan untuk mengungkap topik tersembunyi dalam sekelompok dokumen. Penelitian ini bertujuan untuk melakukan pemodelan topik pada *tweet* berbahasa Indonesia dengan menggunakan *BERTopic*. Proses Pemodelan Topik menggunakan *BERTopic* meliputi langkah-langkah seperti penyematan dokumen, pengurangan dimensi dengan UMAP, pengelompokan dokumen menggunakan HDBSCAN, dan merepresentasikan topik menggunakan c-TF-IDF. Dataset yang digunakan terdiri dari 10.000 *tweet* berbahasa Indonesia yang diambil dari akun Twitter @detikcom. Dari 10.000 *tweet* tersebut, ditemukan 119 topik utama dan 1 topik *outlier*. Evaluasi Pemodelan Topik dilakukan menggunakan *coherence score cv*, dengan hasil rata-rata *coherence score cv* sebesar 0,685, *coherence score cv* tertinggi sebesar 0,995, dan *coherence score cv* terendah sebesar 0,119.

Kata Kunci: Pemodelan Topik, *BERTopic*, *Cohrence Score cv*, *Tweet*

KATA PENGANTAR

Puji syukur kepada Allah SWT atas rahmat dan nikmat-Nya yang telah diberikan kepada penulis sehingga dapat menyelesaikan skripsi ini dengan baik. Skripsi ini disusun sebagai salah satu syarat menyelesaikan pendidikan program Strata-1 di Fakultas Ilmu Komputer Universitas Sriwijaya. Dalam menyelesaikan skripsi ini, penulis menerima bantuan, bimbingan dan dukungan dari banyak pihak, baik secara langsung maupun tidak langsung. Oleh karena itu, penulis ingin menyampaikan terima kasih kepada:

1. Allah SWT atas rahmat dan nikmat-Nya penulis dapat menyelesaikan skripsi ini dengan baik.
2. Kedua orang tua, saudara dan teman yang telah mendoakan, memberi semangat, motivasi, dan nasihat untuk menyelesaikan skripsi ini.
3. Ibu Alvi Syahrini Utami, M.Kom. selaku Ketua Jurusan Teknik Informatika Universitas Sriwijaya.
4. Bapak Dr. Abdiansah, S.Kom., M.Cs. selaku Dosen Pembimbing yang telah membimbing, memberikan motivasi serta arahan kepada penulis dalam proses pengerjaan skripsi.
5. Ibu Novi Yusliani, M.T. selaku Dosen Penguji Tugas Akhir yang telah memberikan ilmu, nasihat serta saran yang membangun.
6. Bapak Rifkie Primartha, M.T. selaku Pembimbing Akademik selama di Universitas Sriwijaya.
7. Seluruh dosen, staf dan pegawai Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
8. Pihak-pihak lain yang tidak dapat penulis sebutkan satu-persatu.

Penulis menyadari bahwa dalam penyusunan skripsi ini masih banyak sekali kekurangan dikarenakan kurangnya pengalaman dan pengetahuan penulis. Oleh karena itu penulis mengharapkan saran dan kritik yang membangun guna kemajuan penelitian selanjutnya. Semoga tugas akhir ini dapat bermanfaat. Terima kasih.

Indralaya, 14 Agustus 2023

Fahmi Guntara Diyasa

DAFTAR ISI

LEMBAR PENGESAHAN SKRIPSI.....	i
TANDA LULUS UJIAN KOMPREHENSIF.....	ii
HALAMAN PERNYATAAN.....	iii
MOTTO DAN PERSEMBAHAN.....	iv
ABSTRACT.....	v
ABSTRAK.....	vi
KATA PENGANTAR.....	vii
DAFTAR ISI.....	ix
DAFTAR TABEL.....	xii
DAFTAR GAMBAR.....	xiii
BAB I PENDAHULUAN.....	I-1
1.1 Pendahuluan.....	I-1
1.2 Latar Belakang.....	I-1
1.3 Rumusan Masalah.....	I-3
1.4 Tujuan Penelitian.....	I-3
1.5 Manfaat Penelitian.....	I-4
1.6 Batasan Masalah.....	I-4
1.7 Sistematika Penulisan.....	I-4
1.8 Kesimpulan.....	I-5
BAB II KAJIAN LITERATUR.....	II-1
2.1 Pendahuluan.....	II-1
2.2 Landasan Teori.....	II-1
2.2.1 Pemodelan Topik.....	II-1
2.2.2 Pra-Pengolahan.....	II-2
2.2.3 <i>BERTopic</i>	II-3
2.2.3.1 <i>Document embeddings</i>	II-3
2.2.3.2 <i>Document clustering</i>	II-3
2.2.3.3 <i>Topic Representation</i>	II-5
2.2.4 Pengukuran Hasil Pemodelan Topik.....	II-6

2.2.5 <i>Rational Unified Process</i>	II-7
2.2.6 Twitter.....	II-8
2.2 Penelitian Lain yang Relevan.....	II-9
2.3 Kesimpulan.....	II-10
BAB III METODOLOGI PENELITIAN.....	III-1
3.1 Pendahuluan.....	III-1
3.2 Pengumpulan Data.....	III-1
3.2.1 Jenis dan Sumber Data.....	III-1
3.2.2 Metode Pengumpulan Data.....	III-1
3.3 Tahapan Penelitian.....	III-2
3.3.1 Menentukan Kerangka Kerja Penelitian.....	III-3
3.3.2 Menentukan Kriteria Pengujian.....	III-4
3.3.3 Menentukan Format Data Pengujian.....	III-5
3.3.4 Menentukan Alat Bantu Penelitian.....	III-5
3.3.5 Melakukan Pengujian Penelitian.....	III-5
3.3.6 Melakukan Analisis dan Menarik Kesimpulan Penelitian.....	III-6
3.4 Metode Pengembangan Perangkat Lunak.....	III-6
3.4.1 Fase Insepsi.....	III-6
3.4.2 Fase Elaborasi.....	III-7
3.4.3 Fase Konstruksi.....	III-7
3.4.4 Fase Transisi.....	III-7
3.5 Kesimpulan.....	III-8
BAB IV METODOLOGI PENELITIAN.....	IV-1
4.1 Pendahuluan.....	IV-1
4.2 Fase Insepsi.....	IV-1
4.2.1 Pemodelan Bisnis.....	IV-1
4.2.2 Kebutuhan Sistem.....	IV-2
4.2.3 Analisis dan Perancangan.....	IV-3
4.2.3.1 Analisis Kebutuhan Perangkat Lunak.....	IV-4
4.2.3.2 Analisis Pra-Pengolahan Data.....	IV-4
4.2.3.3 Analisis Proses Pemodelan Topik.....	IV-7
4.2.3.4 Analisis Hasil Pemodelan Topik.....	IV-10

4.2.4 Implementasi.....	IV-10
4.3 Fase Elaborasi.....	IV-13
4.3.1 Pemodelan Bisnis.....	IV-13
4.3.1.1 Perancangan Data.....	IV-13
4.3.1.2 Perancangan Antarmuka.....	IV-14
4.3.2 Kebutuhan.....	IV-15
4.3.3 Analisis dan Perancangan.....	IV-15
4.3.3.1 Diagram Aktivitas.....	IV-15
4.3.3.2 Diagram Alur.....	IV-17
4.4 Fase Konstruksi.....	IV-18
4.4.1 Kebutuhan.....	IV-19
4.4.2 Implementasi.....	IV-19
4.4.2.1 Implementasi Kelas.....	IV-20
4.4.2.2 Implementasi Antarmuka.....	IV-20
4.5 Fase Transisi.....	IV-21
4.5.1 Pemodelan Bisnis.....	IV-22
4.5.2 Kebutuhan.....	IV-22
4.5.3 Analisis dan Perancangan.....	IV-22
4.5.3.1 Rencana Pengujian.....	IV-22
4.5.3.2 Implementasi.....	IV-24
4.6 Kesimpulan.....	IV-26
BAB V HASIL DAN ANALISIS.....	V-1
5.1 Pendahuluan.....	V-1
5.2 Hasil Penelitian.....	V-1
5.3 Analisis Hasil Penelitian.....	V-3
5.4 Kesimpulan.....	V-4
BAB VI KESIMPULAN DAN SARAN.....	VI-1
6.1 Pendahuluan.....	VI-1
6.2 Kesimpulan.....	VI-1
6.3 Saran.....	VI-1
DAFTAR PUSTAKA.....	xiv

DAFTAR TABEL

Tabel III-1. Contoh tweet yang dikumpulkan.....	III-2
Tabel III-2. Hasil Pengujian.....	III-5
Tabel III-3. Daftar Kata Probabilitas Tertinggi.....	III-6
Tabel IV-1. Kebutuhan Fungsional Perangkat Lunak Pelatihan.....	IV-3
Tabel IV-2. Kebutuhan Non-Fungsional Perangkat Lunak Pelatihan.....	IV-3
Tabel IV-3. Data Tweet.....	IV-4
Tabel IV-4. Data Tweet Setelah Dilakukan Proses <i>Case Folding</i>	IV-5
Tabel IV-5. Data Tweet Setelah Dilakukan Proses <i>Cleaning</i>	IV-6
Tabel IV-6. Tabel Definisi Aktor.....	IV-11
Tabel IV-7. Definisi Use Case.....	IV-11
Tabel IV-8. Skenario Use Case Melakukan Proses <i>Pre-processing</i> Data Masukan.....	IV-11
Tabel IV-9. Skenario Use Case Melakukan Pemodelan Topik Menggunakan BERTopic.....	IV-12
Tabel IV-10. Keterangan Implementasi Kelas.....	IV-20
Tabel IV-11. Rencana Pengujian Use Case Proses Pra-Pengolahan Data.....	IV-23
Tabel IV-12. Rencana Pengujian Use Case Proses Pemodelan Topik Menggunakan BERTopic.....	IV-23
Tabel IV-13. Pengujian Use Case Proses Pra-Pengolahan Data.....	IV-24
Tabel IV-14. Pengujian Use Case Proses Pemodelan Topik Menggunakan BERTopic.....	IV-25
Tabel V-1. Hasil Pemodelan Topik Menggunakan BERTopic.....	V-1
Tabel V-2. Hasil Evaluasi Pemodelan Topik Menggunakan BERTopic.....	V-2

DAFTAR GAMBAR

Gambar II-1. Contoh Case Folding.....	II-2
Gambar II-2. Contoh Cleaning Text.....	II-2
Gambar II-3. Contoh <i>Tokenizing</i>	II-2
Gambar II-4. Tahapan <i>Rational Unified Process</i>	II-7
Gambar III-1. Diagram Tahapan Penelitian.....	III-3
Gambar III-2. Diagram Alur Proses Umum Perangkat Lunak.....	III-3
Gambar IV-1. <i>Output</i> Proses <i>Document Embedding</i> Menggunakan SBERT.....	IV-7
Gambar IV-2. <i>Output</i> Proses <i>Dimensionality Reduction</i> Menggunakan UMAP.....	IV-8
Gambar IV-3. <i>Output</i> Proses <i>Document Clustering</i> Menggunakan HDBSCAN.....	IV-9
Gambar IV-4. <i>Output</i> Proses <i>Topic Represntation</i> Menggunakan <i>c-TF-IDF</i>	IV-9
Gambar IV-5. <i>Use Case</i> Pemodelan Topik Menggunakan <i>BERTopic</i>	IV-10
Gambar IV-6. Rancangan Antarmuka Pra-Pengolahan Data.....	IV-14
Gambar IV-7. Rancangan Antarmuka Hasil Pemodelan Topik.....	IV-14
Gambar IV-8. Diagram Aktivitas Melakukan Pra-Pengolahan Data Pada Sistem	IV-16
Gambar IV-9. Diagram Aktivitas Melakukan Pemodelan Topik.....	IV-16
Gambar IV-10 .Diagram Alur Proses Pra-Pengolahan Data.....	IV-17
Gambar IV-11. Diagram Alur Proses Pemodelan Topik Menggunakan BERTopic.....	IV-18
Gambar IV-12. Diagram Kelas Perangkat Lunak.....	IV-19
Gambar IV-14. Implementasi Antarmuka Pra-Pengolahan Data.....	IV-21
Gambar IV-15. Implementasi Antamuka Hasil Pemodelan Topik.....	IV-21

BAB I

PENDAHULUAN

1.1 Pendahuluan

Pada Bab ini membahas latar belakang masalah, rumusan masalah, tujuan dan manfaat penelitian serta batasan masalah. Bab ini memberikan penjelasan umum mengenai keseluruhan penelitian.

1.2 Latar Belakang

Meningkatnya penggunaan teknologi informasi dan komunikasi dalam beberapa tahun terakhir telah memunculkan satu tren di kalangan masyarakat untuk berkomunikasi ataupun menyampaikan aspirasi melalui media sosial. Salah satu media sosial yang banyak digunakan adalah twitter (Chilmi, 2021). Setiap *tweet* pada twitter memiliki topik tertentu, dan untuk mengetahui topik utama dari kumpulan *tweet* tersebut dapat menggunakan *topic modeling* (Patmawati dan Yusuf, 2021).

Pemodelan topik merupakan teknik yang digunakan dalam pendekatan *text mining* dan *text analysis* dalam menemukan data teks yang tersembunyi dan hubungan antar teks yang saling berkaitan dari suatu korpus (Jelodar et al., 2018). Pemodelan topik telah dibahas pada banyak literatur menggunakan metode-metode yang bervariasi seperti Top2Vec (Angelov, 2020), *Non-negative Matrix Factorization* (NMF) (Carbonetto et al., 2022) dan *Latent Dirichlet*

Allocation (LDA) (Blei et al., 2013). LDA merupakan metode pemodelan topik yang paling populer dan banyak digunakan (Angelov, 2020).

LDA sangat cocok untuk tugas pemodelan topik umum menggunakan berbagai data. Akan tetapi LDA tidak mampu memodelkan hubungan data yang lebih maju dan berkinerja buruk ketika dokumen tidak cukup panjang (Vayansky and Kumar, 2020). NMF adalah pendekatan tanpa pengawasan untuk mengurangi dimensi matriks nonnegatif (Lee and Seung, 1999), dan telah banyak digunakan untuk menemukan hubungan yang mendasari antara teks dan mengidentifikasi topik *laten* (Arora et al, 2012).

Meskipun demikian LDA dan NMF membutuhkan upaya yang cukup besar untuk penyetelan *hyperparameter* guna menciptakan topik yang bermakna (Abuzayed and Al-Khalifa, 2021). Top2Vec dapat digunakan ketika banyak bahasa muncul dalam sebuah korpus (Hendry et al., 2021). Top2Vec juga memungkinkan untuk tidak menggunakan *preprocessing* pada data asli karena telah menggunakan teknik *embeddings* (Egger and Yu, 2022). Akan tetapi Top2Vec tidak mampu bekerja baik dengan data yang kecil misalnya kurang dari 1000 dokumen (Abuzayed and Al-Khalifa, 2021).

BERTopic adalah teknik pemodelan topik yang memanfaatkan penyematan BERT dan TF-IDF berbasis kelas untuk menghasilkan representasi topik yang koheren, juga menggunakan teknik *Uniform Manifold Approximation and Projection* (UMAP) untuk menurunkan dimensi penyematan sebelum pengelompokan dokumen-dokumen (Grootendorst, 2022). *BERTopic* bekerja secara luar biasa dengan penyematan yang telah dilatih sebelumnya dan juga

karena pemisahan antara pengelompokan dokumen dan penggunaan *c-TF-IDF* untuk mengekstraksi representasi topik (Franco and Moreno, 2022).

Karena penggunaan prosedur *c-TF-IDF*, *BERTopic* dapat mendukung beberapa variasi pemodelan topik, seperti pemodelan topik terpadu, pemodelan topik dinamis, atau pemodelan topik berbasis kelas (Abuzayed and Al-Khalifa, 2021). Kelebihan utamanya terletak pada kenyataan bahwa algoritma ini bekerja dengan baik pada sebagian besar aspek domain pemodelan topik, sedangkan yang lain biasanya unggul dalam satu aspek (Egger and Yu, 2022). Berdasarkan referensi penelitian yang dilakukan sebelumnya, metode *BERTopic* akan digunakan dalam penelitian pemodelan topik pada *tweet* bahasa Indonesia.

1.3 Rumusan Masalah

Berdasarkan permasalahan yang telah dijelaskan pada latar belakang maka rumusan masalah dari penelitian ini adalah

1. Bagaimana mengembangkan pemodelan topik pada *tweet* bahasa Indonesia menggunakan *BERTopic*?
2. Bagaimana kinerja topik yang dihasilkan dari pemodelan topik pada *tweet* bahasa Indonesia menggunakan *BERTopic*?

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah :

1. Menghasilkan pemodelan topik pada *tweet* bahasa Indonesia menggunakan *BERTopic*.

2. Mengetahui hasil pemodelan topik pada *tweet* bahasa Indonesia menggunakan *BERTopic*.

1.5 Manfaat Penelitian

Manfaat yang diperoleh dari penelitian ini adalah :

1. Membantu pembaca untuk mengetahui topik *laten* pada kumpulan *tweet* berbahasa Indonesia.
2. Hasil penelitian dapat dijadikan sebagai rujukan penelitian terkait.

1.6 Batasan Masalah

Batasan masalah dari penelitian ini adalah :

1. Data yang digunakan adalah data *tweet* berbahasa Indonesia.
2. Data *tweet* yang digunakan berjumlah 10.000 *tweet*.

1.7 Sistematika Penulisan

Sistematika penulisan tugas akhir mengikuti standar penulisan tugas akhir Fakultas Ilmu Komputer Universitas Sriwijaya yaitu sebagai berikut:

BAB I. PENDAHULUAN

Pada bab ini membahas latar belakang, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah serta sistematika penelitian yang akan dijadikan sebagai pokok pikiran penelitian ini.

BAB II. KAJIAN LITERATUR

Pada bab ini membahas landasan teori yang digunakan dalam penelitian, seperti definisi pemodelan topik dan model *BERTopic*, serta beberapa literatur yang relevan dengan penelitian ini.

BAB III. METODOLOGI PENELITIAN

Pada bab ini membahas proses yang akan dilaksanakan selama penelitian, Seperti pengumpulan data, analisis data dan perancangan perangkat lunak. Setiap tahap akan dijelaskan berdasarkan kerangka kerja yang dibuat.

BAB IV. PENGEMBANGAN PERANGKAT LUNAK

Pada bab ini membahas analisis dan rancangan perangkat lunak yang akan dikembangkan. Diawali dari analisis kebutuhan, perancangan dan konstruksi perangkat lunak, dan diakhiri dengan evaluasi untuk memastikan sistem yang dikembangkan sudah sesuai dengan rancangan dan kebutuhan penelitian.

BAB V. HASIL DAN ANALISIS PENELITIAN

Pada bab ini menyajikan hasil pengujian berdasarkan langkah-langkah yang telah direncanakan. Analisis diberikan sebagai dasar kesimpulan yang akan diambil dari penelitian ini.

BAB VI. KESIMPULAN DAN SARAN

Pada bab ini membahas kesimpulan yang diambil berdasarkan uraian dalam bab sebelumnya serta saran yang diberikan berdasarkan penelitian yang telah dilakukan.

1.8 Kesimpulan

Pada bab ini telah dijelaskan latar belakang, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah serta sistematika penelitian yang akan dijadikan sebagai pokok pikiran penelitian ini.

BAB II

KAJIAN LITERATUR

2.1 Pendahuluan

Pada bab ini akan dijelaskan teori yang melandasi penelitian. Bab ini berisi penjelasan mengenai pemodelan topik, twitter, *BERTopic*, beberapa penelitian yang relevan, serta kesimpulan.

2.2 Landasan Teori

2.2.1 Pemodelan Topik

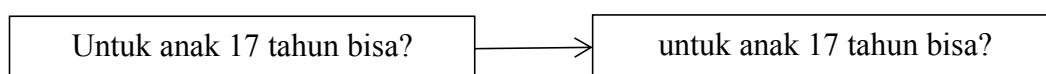
Pemodelan topik merupakan algoritma yang bertujuan untuk menemukan topik yang tersembunyi dari rangkaian kata dalam dokumen yang tidak terstruktur (Putra, 2017). Algoritma *Topic modeling* menganalisis kata-kata dari teks asli untuk menemukan topik yang berada diantara teks tersebut, bagaimana topik topik saling terhubung satu sama lain, dan bagaimana tema-tema tersebut dapat berubah dari waktu ke waktu, sehingga dapat dikembangkan untuk pencarian, ataupun meringkas teks yang terdapat dalam dokumen (Blei et al., 2003).

Pendapat lain terkait *Topic modeling* disampaikan oleh Meeks and Weingart (2012), yang menyatakan bahwa *Topic modeling* merupakan bentuk text mining, sebagai salah satu metode untuk mengidentifikasi pola dalam sebuah korpus. Topic modeling dapat pula dikatakan sebagai sebuah tool untuk mengubah korpus yang berbentuk kumpulan kata, menjadi topik yang dapat menggambarkan korpus tersebut.

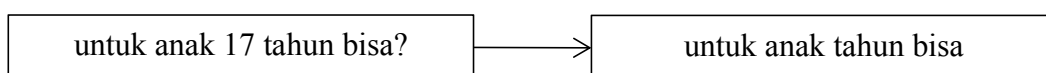
Algoritma pemodelan topik tidak memerlukan penjelasan sebelumnya atau pelabelan karena topik dokumen muncul dari analisis teks asli. Pada umumnya, *Topic modeling* diaplikasikan pada jumlah dokumen yang sangat besar dan dapat mengelola dokumen-dokumen individual sesuai topik yang ditemukan, sehingga *Topic modeling* memungkinkan kita untuk mengatur dan meringkas arsip elektronik pada skala yang tidak mungkin dengan penjelasan manusia secara manual (Blei et al., 2003).

2.2.2 Pra-Pengolahan

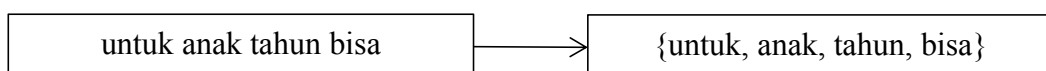
Pra-pengolahan teks merupakan langkah yang dilakukan untuk mempersiapkan teks sehingga menjadi lebih terstruktur. Contoh tahapan yang dapat dilakukan pada proses pra-pengolahan teks adalah *case folding*, *cleaning text* dan *tokenizing* (Pradha et al., 2019). *Case folding* adalah proses menyamakan karakter huruf dalam teks menjadi huruf kecil atau huruf besar. *Cleaning text* adalah proses pembersihan teks dari karakter-karakter yang tidak diinginkan. *Tokenizing* adalah proses memisahkan teks menjadi kepingan-kepingan kata.



Gambar II-1. Contoh Case Folding



Gambar II-2. Contoh Cleaning Text



Gambar II-3. Contoh *Tokenizing*

2.2.3 *BERTopic*

BERTopic adalah metode pemodelan topik yang memanfaatkan *embedding* BERT dan *c-TF-IDF* untuk membuat *dense cluster* yang memungkinkan topik dengan mudah ditafsirkan sambil menyimpan kata-kata penting dalam deskripsi topik (Grootendorst, 2022). Dalam melakukan pemodelan topik, metode *BERTopic* memiliki tiga tahapan yaitu melakukan *document embedding*, *document clustering* untuk melakukan *cluster* ke dalam bentuk *semantic similar cluster*, lalu *topic representation* untuk membuat representasi topik dari masing-masing *cluster*.

2.2.3.1 *Document embeddings*

Dalam prosedur *BERTopic* yang akan digunakan dalam penelitian ini, peneliti akan menggunakan model *Sentence-BERT* (SBERT) untuk melakukan *document embedding*. SBERT merupakan modifikasi dari jaringan *pretrained* BERT yang menggunakan struktur jaringan *siamese* sehingga dapat memperoleh *sentence embedding* yang bermakna secara semantik dan dapat dibandingkan menggunakan *cosine-similarity* (Reimers and Gurevych, 2019). SBERT mencapai performa yang luar biasa dalam beragam tugas penyisipan kalimat (Thakur et al., 2020).

2.2.3.2 *Document clustering*

Sebelum melakukan *clustering*, dimensi dari penyematan yang dihasilkan dikurangi untuk mengoptimalkan proses *clustering*. Proses ini dilakukan karena hasil dari *document embedding* akan meningkatkan dimensi data, sehingga perlu

dilakukan *dimensionality reduction* (Grootendorst, 2022), yaitu merubah data vektor yang memiliki dimensi tinggi menjadi dimensi yang lebih rendah sehingga data tersebut bisa diproses oleh algoritma *clustering*.

UMAP (*Uniform Manifold Approximation and Projection*) adalah teknik terbaru pembelajaran *manifold* untuk reduksi dimensi (McInnes et al., 2018). UMAP bekerja dengan cara menemukan representasi rendah dimensi dari data asli, sehingga memungkinkan visualisasi data yang tinggi atau *multidimensional* menjadi lebih mudah dipahami. UMAP didasarkan pada teori geometri dan topologi aljabar *Riemannian* untuk memahami struktur data dalam ruang fitur yang tinggi.

UMAP mampu mempertahankan struktur global data dengan lebih baik daripada t-SNE, dan menawarkan performa waktu yang lebih cepat, terutama ketika berurusan dengan dataset yang lebih besar. Selain itu, (Allaoui et al., 2020) menunjukkan bahwa mengurangi penyematan dimensi tinggi dengan UMAP dapat meningkatkan kinerja algoritma *clustering* yang terkenal, seperti k-Means dan HDBSCAN, baik dalam hal akurasi pengelompokan maupun waktu. Oleh karena itu, penelitian ini menggunakan UMAP untuk mengurangi dimensi penyematan dokumen yang dihasilkan.

Proses *clustering* dari hasil *document embedding* yang telah melalui proses *dimensionality reduction* dilakukan dengan metode HDBSCAN. HDBSCAN (*Hierarchical Density Based Spatial Clustering of Application with Noise*) adalah perluasan dari DBSCAN yang menemukan klaster dengan kepadatan yang bervariasi dengan mengubah DBSCAN menjadi algoritma pengelompokan hierarkis (McInnes and Healy, 2017). HDBSCAN memodelkan *cluster*

menggunakan pendekatan *soft-clustering* yang memungkinkan *noise* dimodelkan sebagai *outlier*. HBDSCAN mencegah dokumen yang tidak terkait ditugaskan ke *cluster* mana pun dan diharapkan dapat meningkatkan representasi topik.

2.2.3.3 Topic Representation

Representasi topik dimodelkan berdasarkan dokumen di setiap *cluster* dimana setiap *cluster* akan diberi satu topik. *TF-IDF* berbasis kelas digunakan untuk membuat representasi topik pada setiap *cluster*.

Prosedur *TF-IDF* klasik mengkombinasikan dua statistik, *term frequency*, dan *inverse document frequency* (Joachims, 1996):

$$W_{t,d} = \text{tf}_{t,d} \log \left(\frac{N}{\text{df}_t} \right) \quad (\text{II-1})$$

Di mana *term frequency* memodelkan frekuensi kata t dalam dokumen d dan *inverse document frequency* mengukur berapa banyak informasi suatu istilah atau kata tersedia untuk dokumen dan dihitung dengan mengambil logaritma dari jumlah dokumen di sebuah korpus N dibagi dengan jumlah dokumen yang mengandung t .

Kemudian prosedur ini di generalisasikan ke dalam *cluster* dokumen. Pertama, semua dokumen dalam sebuah *cluster* diperlakukan sebagai satu dokumen hanya dengan menggabungkan dokumen. Kemudian, *TF-IDF* disesuaikan untuk representasi ini dengan menerjemahkan dokumen ke *cluster* :

$$W_{t,c} = \text{tf}_{t,c} \log \left(1 + \frac{A}{\text{tf}_t} \right) \quad (\text{II-2})$$

Dimana *term frequency* memodelkan frekuensi kata t di kelas c atau dalam contoh ini. Di Sini, kelas c adalah kumpulan dokumen yang digabungkan menjadi

satu dokumen dalam *cluster*. Kemudian, *inverse document frequency* diganti dengan *inverse class frequency* untuk mengukur berapa banyak informasi yang disediakan sebuah kata untuk suatu kelas. Itu dihitung dengan mengambil logaritma dari jumlah rata-rata kata per kelas A dibagi dengan frekuensi kata t di semua kelas. Untuk hanya menampilkan nilai positif, kami menambahkan satu ke pembagian dalam logaritma.

Dengan demikian, prosedur *TF-IDF* berbasis kelas ini memodelkan pentingnya kata-kata dalam kelompok, bukan dokumen individu. Hal ini memungkinkan kita untuk menghasilkan distribusi topik kata untuk setiap kelompok dokumen. Terakhir, dengan menggabungkan perwakilan *c-TF-IDF* secara iteratif dari topik yang paling tidak umum dengan topik yang paling umum. Dengan demikian, kita dapat mengurangi jumlah topik menjadi nilai yang ditentukan oleh pengguna.

2.2.4 Pengukuran Hasil Pemodelan Topik

Pengukuran *coherence score* terhadap sebuah topik dilakukan dengan mengukur derajat kemiripan semantik antar kata yang memiliki skor atau probabilitas tertinggi pada suatu topik. Dalam pengukuran *coherence score*, terdapat beberapa metode pengukuran salah satunya adalah *coherence score cv*. *Coherence score cv* mengukur kesamaan berdasarkan perhitungan *cosine similarity* antara kata-kata dalam suatu topik. Untuk menghitungnya, terlebih dahulu, dibuat suatu representasi vektor untuk setiap kata berdasarkan distribusi kata tersebut dalam korpus. Kemudian, dihitung kesamaan kosinus antara vektor kata-kata dalam topik. Skor kesamaan dihitung sebagai rata-rata dari kesamaan

kosinus antara semua pasangan kata dalam topik. Pengukuran *coherence score cv* digunakan persamaan II-3.

$$Coherence_{cv} = \frac{2}{W(W-1)} \sum_{i=1}^W \sum_{j=i+1}^W Cosine_Similarity(w_i, w_j) \quad (II-3)$$

W adalah jumlah kata dalam topik

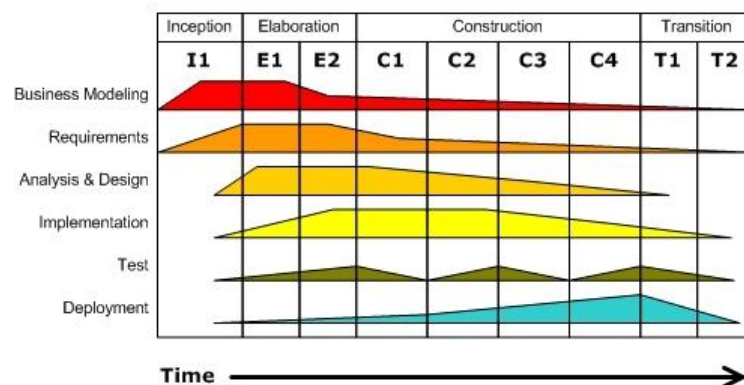
w_i dan w_j adalah vektor

$Cosine_Similarity(w_i, w_j)$ adalah fungsi yang menghitung kesamaan kosinus antara dua vektor w_i dan w_j

2.2.5 Rational Unified Process

Rational Unified Process (RUP) adalah proses pengembangan perangkat lunak yang memberikan pendekatan disiplin untuk menetapkan tugas dan tanggung jawab dalam organisasi pembangunan. RUP bertujuan untuk menghasilkan perangkat lunak yang berkualitas tinggi dan dapat memenuhi kebutuhan pengguna akhir (Gornik, 2003).

RUP menerapkan pengembangan perangkat lunak berorientasi objek dengan berfokus pada model pengembangan menggunakan *Unified Modeling Language* (UML). Tahapan *Rational Unified Process* (RUP) ditunjukkan pada Gambar II-5.



Gambar II-4. Tahapan *Rational Unified Process*

Gambar II-5 menunjukkan tahapan RUP dalam dua dimensi. Dimensi pertama yang ditunjukkan sumbu vertikal merepresentasikan aspek statis proses, dinyatakan dalam aktivitas, pekerja dan alur kerja. Dimensi kedua yang ditunjukkan sumbu horizontal mewakili waktu dan menunjukkan aspek dinamis proses, dinyatakan dalam siklus, fase, iterasi dan tonggak. Satu siklus pengembangan dalam RUP dibagi menjadi empat fase yaitu inepsi, elaborasi, konstruksi dan transisi.

1. Fase inepsi, pada fase awal RUP ini *business case* ditetapkan dan ruang lingkup proyek dibatasi. *Business case* mencakup kriteria keberhasilan, penilaian resiko, perkiraan sumber daya yang diperlukan, dan rencana fase pengembangan.
2. Fase elaborasi, fase ini bertujuan untuk menganalisis domain masalah, membangun fondasi arsitektur yang kuat, mengembangkan rencana proyek dan menghilangkan elemen resiko terbesar proyek.
3. Fase konstruksi, selama fase konstruksi, semua komponen dan fitur aplikasi dikembangkan dan diintegrasikan ke dalam produk, kemudian semua fitur diuji secara menyeluruh.
4. Fase transisi, tujuan dari fase transisi adalah untuk mentransisikan produk perangkat lunak ke komunitas pengguna. Selama fase ini, akan dilakukan pengujian beta, pelatihan pengguna dan pengelola dan meluncurkan produk.

2.2.6 Twitter

Twitter adalah situs web dimiliki dan dioperasikan oleh Twitter, Inc. yang menawarkan jaringan sosial berupa *microblog*. Disebut *microblog* karena situs ini memungkinkan pengguna mengirim dan membaca pesan blog seperti pada

umumnya namun terbatas hanya sejumlah 140 karakter yang ditampilkan pada halaman profil pengguna. Twitter memiliki karakteristik dan format penulisan yang unik dengan simbol ataupun aturan khusus. Pesan dalam Twitter dikenal dengan sebutan *tweet*.

2.2 Penelitian Lain yang Relevan

Penelitian-penelitian yang relevan dapat dijadikan sebagai sumber referensi peneliti dalam melakukan penelitian serta untuk menguatkan kerangka berfikir dalam menyelesaikan masalah yang diteliti. Penelitian-penelitian tersebut diambil dari berbagai sumber ilmiah baik prosiding maupun jurnal ilmiah.

Putra (2017) melakukan penelitian analisis topik informasi publik media sosial di Surabaya menggunakan pemodelan *Latent Dirichlet Allocation* (LDA). Dengan metode LDA didapatkan jumlah topik yang terdapat dalam pesan media sosial yaitu 4 topik dengan nilai *perplexity* terbaik yaitu sebesar 213.41.

Al-Khairi et al (2019) melakukan penelitian pendetektisian topik fashion di Twitter dengan mengimplementasikan LDA dan *Gibbs Sampling*. Berdasarkan hasil penelitian, konfigurasi parameter 20 topik dengan 1.000 iterasi memperoleh skor UMass terbaik dengan nilai -56.342, dan konfigurasi parameter 50 topik dengan 1.000 iterasi memperoleh skor PMI atau UCI terbaik dengan nilai 6.272. Kesimpulan yang didapat pada penelitian tersebut adalah metode LDA dapat menghasilkan topik mengenai fashion yang sedang dibicarakan di media sosial Twitter. Meskipun tidak semua topik yang dihasilkan memberikan hasil yang diinginkan.

Egger and Yu (2022) melakukan penelitian perbandingan topik antara LDA, NFM, Top2Vec dan *BERTopic*. Dari penelitian ini diperoleh bahwa *BERTopic* bekerja secara luar biasa dengan penyematan yang telah dilatih sebelumnya. Karena prosedur *c-TF-IDF*, *BERTopic* dapat mendukung beberapa variasi pemodelan topik, seperti pemodelan topik terpandu, pemodelan topik dinamis, atau pemodelan topik berbasis kelas. Kekuatan utamanya terletak pada kenyataan bahwa algoritme bekerja dengan baik pada sebagian besar aspek domain pemodelan topik, sedangkan yang lain biasanya unggul dalam satu aspek.

Top2Vec dapat menskalakan sejumlah besar topik dan sejumlah besar data. Kelebihan seperti ini sangat dibutuhkan ketika banyak bahasa muncul dalam sebuah korpus. Akan tetapi kelemahan Top2Vec tidak memenuhi syarat untuk bekerja dengan sejumlah kecil data misalnya <1.000 dokumen. LDA dan NMF biasanya memerlukan asumsi rinci mengenai hyperparameter untuk menemukan jumlah topik yang optimal, dan ini terbukti menjadi tugas yang sulit. Akan tetapi NMF bekerja dengan baik dengan teks yang lebih pendek seperti tweet. Berdasarkan perincian tertentu selama prosedur analitis dan masalah kualitas, penelitian ini menyoroti kemandirian penggunaan *BERTopic* dan NMF untuk menganalisis data twitter.

2.3 Kesimpulan

Pada bab ini telah dijelaskan teori-teori yang berkaitan dengan penelitian. Bab ini juga membahas penelitian-penelitian lain yang relevan dengan pemodelan topik pada teks bahasa Indonesia menggunakan metode *BERTopic*.

BAB III

METODOLOGI PENELITIAN

3.1 Pendahuluan

Pada bab ini akan membahas mengenai metodologi penelitian yang menguraikan rencana tahapan penelitian serta manajemen proyek penelitian. Rencana tahapan penelitian akan digunakan sebagai landasan pada pengembangan perangkat lunak.

3.2 Pengumpulan Data

Pada bagian ini dijelaskan mengenai tahapan pengumpulan data yang akan digunakan dalam penelitian.

3.2.1 Jenis dan Sumber Data

Data yang digunakan sebagai objek penelitian ini adalah jenis data primer. Data berasal dari salah satu akun twitter media masa Indonesia yaitu detik news dengan username twitter *@detikcom*.

3.2.2 Metode Pengumpulan Data

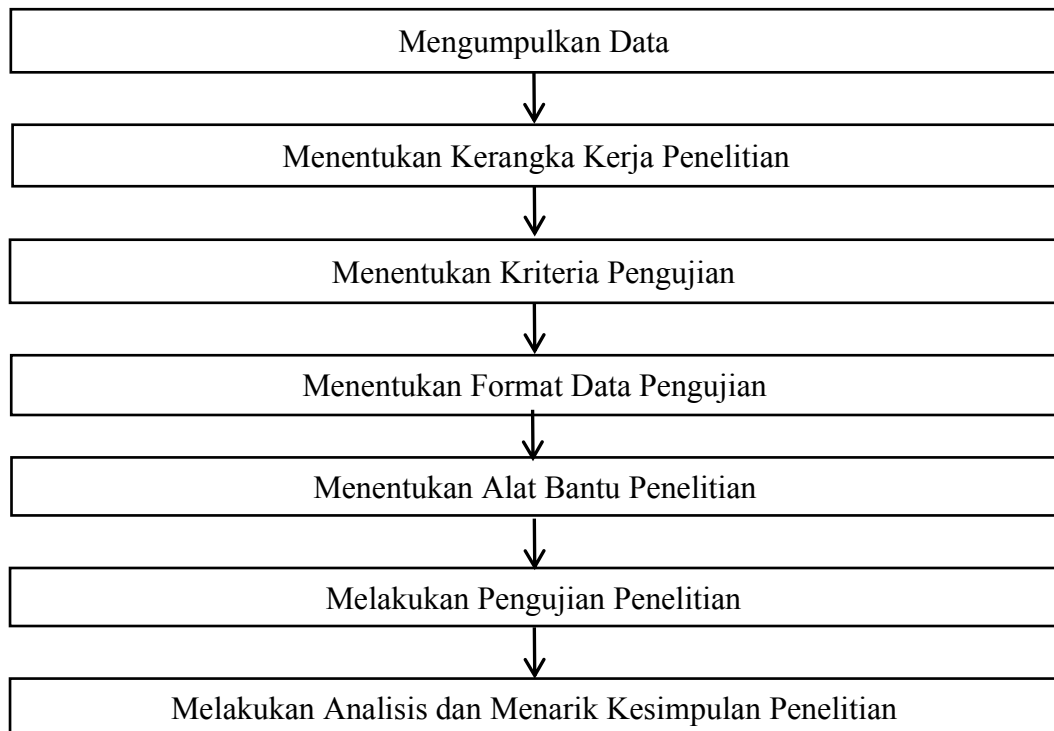
Metode yang dipakai untuk mengumpulkan data pada penelitian ini adalah teknik *scrapping* menggunakan library *snsrape*. Data tweet dikumpulkan dan di export dalam bentuk file csv. Data yang akan digunakan pada penelitian ini berjumlah 10.000 tweet. Tabel III-1 menampilkan contoh data tweet yang telah dikumpulkan.

Tabel III-1. Contoh *tweet* yang dikumpulkan

No	<i>Tweet</i>
1.	Jika ingin ke wisata kuliner terbaik di Bali, maka harus sabar bila susah parkirnya. Ini empat rekomendasinya! https://t.co/vwbKjbC47M
2.	Viral di TikTok, seorang wanita membatalkan pernikahannya karena menemukan folder rahasia di komputer milik tunangannya. Apa isi folder tersebut? https://t.co/HSXfIUFIop
3.	Song Joong Ki kini tak jomblo lagi. Ia dikonfirmasi pacaran dengan wanita dari kalangan non selebriti asal Inggris. Ini sosoknya. https://t.co/yWQAmSL4EE
4.	Presiden Jokowi meresmikan pengembangan Stasiun Manggarai tahap 1 di Stasiun Manggarai hari ini, Senin (26/12/2022). https://t.co/StGeb28iOa
5.	Drama Korea yang dibintangi Song Joong Ki, Reborn Rich tamat. Bersamaan dengan tamatnya Reborn Rich di episode 16, sang aktor mengaku punya pacar. https://t.co/Wf5F3rWyfm

3.3 Tahapan Penelitian

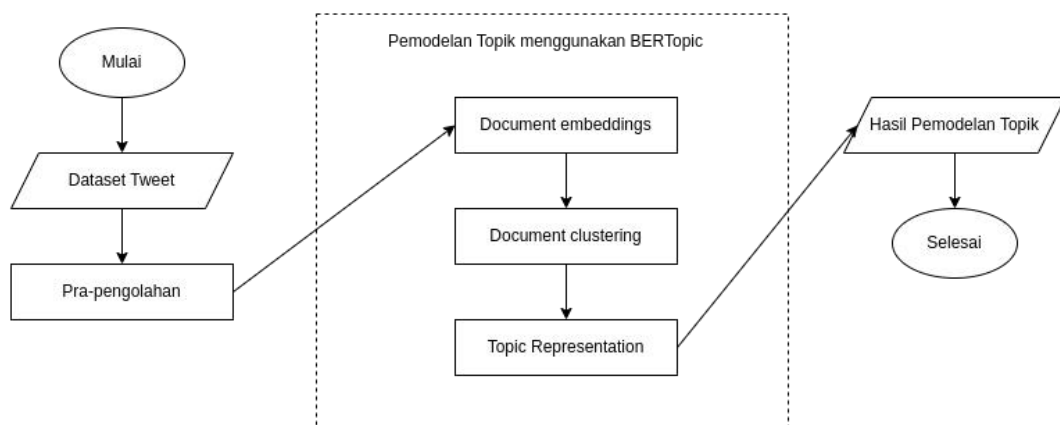
Tahapan penelitian adalah rincian kegiatan yang akan dilakukan selama penelitian berlangsung. Tahapan penelitian ini adalah sebagai berikut:



Gambar III-1. Diagram Tahapan Penelitian

3.3.1 Menentukan Kerangka Kerja Penelitian

Kerangka kerja penelitian ini adalah sebagai berikut:



Gambar III-2. Diagram Alur Proses Umum Perangkat Lunak

Berdasarkan Gambar III-2, kerangka kerja yang akan dilakukan pada penelitian ini adalah sebagai berikut:

a. Pra-pengolahan

Pra-Pengolahan merupakan tahapan pertama yang akan dilakukan terhadap data teks untuk menyiapkan teks dari data tidak terstruktur menjadi data yang terstruktur. Tahapan pra-pengolahan yang dilakukan pada penelitian ini yaitu *cleaning* dan *case folding*. Pada proses *cleaning* data yang memiliki karakter angkat, simbol dan tanda baca akan dihilangkan. Selanjutnya pada proses *case folding* dilakukan penyeragaman kata pada dokumen menjadi huruf kecil, dan hanya karakter ‘a’ hingga ‘z’ yang diterima.

b. Pemodelan Topik menggunakan *BERTopic*

BERTopic menghasilkan representasi topik melalui tiga langkah. Pertama, setiap dokumen dikonversi ke representasi penyematannya menggunakan model bahasa terlatih SBERT. Kemudian, sebelum mengelompokkan penyematan ini, dimensi dari penyematan yang dihasilkan dikurangi untuk mengoptimalkan proses pengelompokan menggunakan UMAP dan HDBSCAN. Terakhir, dari kumpulan dokumen, representasi topik diekstraksi menggunakan *c-TF-IDF*.

3.3.2 Menentukan Kriteria Pengujian

Pengujian dilakukan dengan menggunakan data uji berupa sekelompok dokumen dalam bentuk format file .csv yang telah dikumpulkan dari tweet detikcom dengan menggunakan teknik *scraping*. Pengujian akan dilakukan dengan cara menghitung *coherence score cv* setiap topik.

3.3.3 Menentukan Format Data Pengujian

Format data pengujian pada penelitian ini menggunakan pengukuran *coherence score cv* untuk setiap topik.

Tabel III-2. Hasil Pengujian

Topik	Daftar Kata	<i>Coherence Score cv</i>

3.3.4 Menentukan Alat Bantu Penelitian

Alat bantu yang digunakan dalam proses penelitian pemodelan topik pada *tweet* bahasa Indonesia menggunakan *BERTopic* adalah sebagai berikut.

1. Perangkat Keras

Processor : Amd Rayzen 5

RAM : 8 GB

SSD : 500 GB

2. Perangkat Lunak

Sistem Operasi : Linux Ubuntu 20.04.5 LTS

IDE : Spyder, Google Colab

3.3.5 Melakukan Pengujian Penelitian

Model *BERTopic* yang telah dihasilkan oleh sistem akan diujikan untuk mendapatkan *coherence score cv* untuk setiap topiknya. *Coherence score cv* dapat

diukur dengan membandingkan model *BERTopic* dengan korpus yang sama yang digunakan dalam pemodelan topik. Pengujian dilakukan untuk setiap topik dari hasil pemodelan topik menggunakan *BERTopic*.

3.3.6 Melakukan Analisis dan Menarik Kesimpulan Penelitian

Setelah *coherence score cv* didapatkan untuk setiap topik, maka akan dilakukan analisis terhadap nilai tersebut. Data hasil pengujian ini akan disajikan dalam bentuk tabel. Pada Tabel III-4, dapat dilihat *coherence score cv* untuk masing-masing topik. *Coherence score cv* ini dapat digunakan untuk menentukan kualitas dari topik itu sendiri. Semakin besar *coherence score cv*, maka semakin baik kualitas topiknya.

Tabel III-3. Daftar Kata Probabilitas Tertinggi

Topik	Daftar Kata	<i>Coherence Score cv</i>

3.4 Metode Pengembangan Perangkat Lunak

Metode yang digunakan adalah RUP (*Rational Unified Process*). Dalam RUP proses pengembangan dilakukan dalam empat fase, yaitu inepsi, elaborasi, konstruksi, dan transisi.

3.4.1 Fase Inepsi

Hal-hal yang dilakukan pada fase ini adalah:

1. Pemodelan bisnis: Menentukan ruang lingkup masalah.
2. Kebutuhan: Menentukan semua kebutuhan perangkat lunak.
3. Analisis dan perancangan: Menganalisis alur proses, data yang dikumpulkan serta kebutuhan fungsional dan non-fungsional perangkat lunak yang dibangun.
4. Implementasi: Merancang diagram *use case* berdasarkan kebutuhan yang telah ditentukan.

3.4.2 Fase Elaborasi

Hal-hal yang dilakukan pada fase ini adalah:

1. Pemodelan bisnis: Merancang dan membuat antarmuka perangkat lunak.
2. Kebutuhan: Menentukan kebutuhan perangkat lunak.
3. Analisis dan perancangan: Membuat model diagram aktivitas dan diagram alur.

3.4.3 Fase Konstruksi

Hal-hal yang dilakukan pada fase ini adalah:

1. Kebutuhan: Membuat model diagram kelas perangkat lunak.
2. Implementasi: Membangun program menggunakan bahasa pemrograman yang telah ditentukan.

3.4.4 Fase Transisi

Hal-hal yang dilakukan pada fase ini adalah:

1. Pemodelan bisnis: Menentukan pengujian perangkat lunak.
2. Kebutuhan: Menentukan alat bantu pengujian yang akan digunakan.

3. Analisis dan perancangan: Membuat *use case* pengujian.
4. Implementasi: Melaksanakan pengujian perangkat lunak berdasarkan *use case* yang telah dibuat.

3.5 Kesimpulan

Bab ini telah memberikan rincian tahapan penelitian yang akan dilakukan, yaitu: mengumpulkan data, menentukan kriteria dan format data yang diperlukan untuk proses pengujian serta analisis hasil pengujian. Bab ini juga membahas jadwal pembuatan dan penelitian sistem.

BAB IV

METODOLOGI PENELITIAN

4.1 Pendahuluan

Pada bab 3 telah dijelaskan bahwa penelitian ini memerlukan sebuah alat yaitu perangkat lunak. Oleh sebab itu sebuah perangkat lunak akan dikembangkan menggunakan *Rational Unified Process* (RUP) untuk melakukan pemodelan topik pada *tweet* bahasa Indonesia menggunakan *BERTopic*.

4.2 Fase Insepsi

Tahap pertama dalam pengembangan perangkat lunak adalah melakukan identifikasi kebutuhan sistem berdasarkan permasalahan dilihat dari sisi pengguna. Kegiatan yang akan dilakukan pada fase ini antara lain proses analisis awal sistem, proses identifikasi dan spesifikasi kebutuhan sebelum adanya perangkat lunak serta proses pemodelan diagram *use case*.

4.2.1 Pemodelan Bisnis

Dalam beberapa tahun terakhir, ada sebuah tren baru di masyarakat yang muncul karena peningkatan penggunaan teknologi informasi dan komunikasi, yaitu berkomunikasi atau menyampaikan aspirasi melalui media sosial. Salah satu media sosial yang paling populer digunakan adalah twitter. Setiap tweet di twitter memiliki topik-topik tertentu dan dengan menggunakan teknik *topic modeling*, dapat ditentukan topik utama dari sekumpulan *tweet* tersebut.

Melalui teknik pemodelan topik pada tweet, dapat dilakukan analisis terhadap opini masyarakat mengenai suatu produk, jasa, atau peristiwa tertentu. Selain itu, teknik ini juga dapat digunakan untuk mengidentifikasi dan menganalisis peristiwa yang terjadi di dunia nyata, seperti bencana alam, perkembangan politik, atau kejadian olahraga. Pemodelan topik pada *tweet* juga dapat digunakan untuk memahami perilaku konsumen, menentukan pasar sasaran, dan mengevaluasi kampanye pemasaran.

Perangkat lunak yang dikembangkan pada penelitian ini berbasis desktop dengan menggunakan bahasa pemrograman Python. Perangkat lunak yang dihasilkan bertujuan untuk mengetahui topik *laten* pada sekumpulan tweet bahasan Indonesia dan juga untuk mengetahui kinerja topik yang dihasilkan oleh *BERTopic* dalam melakukan pemodelan topik.

4.2.2 Kebutuhan Sistem

Kebutuhan sistem yang akan disediakan pada perangkat lunak dibagi dengan tiga fitur utama yakni fitur proses data, fitur pemodelan topik dan fitur evaluasi pemodelan topik. Fitur proses data bertujuan untuk melakukan proses pra-pengolahan data teks masukan pengguna. Data tersebut akan diolah untuk memudahkan proses pemodelan topik. Proses pra-pengolahan data ini meliputi case folding dan cleaning.

Fitur pemodelan topik bertujuan untuk menemukan daftar topik yang ada pada sekumpulan dokumen. Pemodelan topik menggunakan *BERTopic* menghasilkan topik melalui tiga langkah. *Document embeddings*, *document clustering* dan *topic representation*. Fitur evaluasi pemodelan topik bertujuan untuk mengetahui kualitas topik dari hasil pemodelan topik.

Dalam merealisasikan fitur-fitur tersebut perangkat lunak harus memenuhi kebutuhan fungsional maupun kebutuhan non-fungsional. Tabel IV-1 dan Tabel IV-2 menunjukkan kebutuhan fungsional dan kebutuhan non-fungsional perangkat lunak pelatihan.

Tabel IV-1. Kebutuhan Fungsional Perangkat Lunak Pelatihan

No	Kebutuhan Fungsional
1	Perangkat lunak dapat melakukan proses pra-pengolahan data masukan.
2	Perangkat lunak dapat melakukan proses pemodelan topik menggunakan <i>BERTopic</i> .
3	Perangkat lunak dapat melakukan proses evaluasi pemodelan topik

Tabel IV-2. Kebutuhan Non-Fungsional Perangkat Lunak Pelatihan

No	Kebutuhan Fungsional
1	Perangkat lunak memiliki <i>Interface</i> yang mudah dipahami dan digunakan oleh pengguna.

4.2.3 Analisis dan Perancangan

Kegiatan yang dilakukan pada tahapan ini adalah analisis kebutuhan perangkat lunak, analisis pra-pengolahan data, analisis proses pemodelan topik, analisis hasil pemodelan topik dan implementasi.

4.2.3.1 Analisis Kebutuhan Perangkat Lunak

Berdasarkan uraian pemodelan bisnis, perangkat lunak harus memiliki beberapa proses kemampuan. Beberapa proses tersebut meliputi proses pra-pengolahan, proses pemodelan topik, dan proses evaluasi pemodelan topik.

4.2.3.2 Analisis Pra-Pengolahan Data

Tabel IV-3. Data *Tweet*

No	Tweet
1	Jadwal Liga Inggris pekan ini akan menyajikan dua derby besar. Akan ada derby Manchester di Old Trafford dan derby London Utara di markas Tottenham Hotspur. https://t.co/63QQt8fx0u
2	Polres Jakbar mengungkap kasus investasi bodong mengatasnamakan waralaba Double Dipps. Total kerugian korban mencapai Rp 19,6 miliar. https://t.co/cx7ZhoRvX1
3	Tiket pesawat harganya diklaim sudah turun saat ini. Benarkah demikian? https://t.co/viM8PMasnV
4	Setelah sebelumnya ada pria Jepang yang ingin hidup sebagai hewan, kini ada lagi yang punya keinginan serupa. Bedanya, dia mau jadi serigala. https://t.co/hry3diecIE
5	Belakangan ini kasus keracunan nitrogen cair dalam jajanan chiki ngebul tengah ramai diperbincangkan masyarakat. Apakah berobatnya bisa menggunakan BPJS? Ini jawabannya! https://t.co/EnmkUvqXzj

Data yang digunakan dalam pengembangan perangkat lunak berasal dari salah satu akun twitter media masa Indonesia yaitu detik news dengan username twitter @detikcom. Data tweet dikumpulkan dan di *export* dalam bentuk file csv. Data yang akan digunakan pada penelitian ini berjumlah 10.000 tweet. Proses pra-pengolahan berfungsi agar data yang digunakan pada proses pemodelan topik menjadi lebih terstruktur dan tidak menimbulkan bias. Tabel IV-3 menunjukkan lima contoh data komentar sebelum melalui proses pra-pengolahan.

Tahapan pra-pengolahan yang dilakukan adalah sebagai berikut:

1. *Case Folding*

Case folding bertujuan untuk menyeragamkan bentuk data awal yang berupa huruf kapital (*uppercase*) menjadi huruf kecil (*lowercase*). Tabel IV-4 menunjukkan hasil *case folding* dari contoh data komentar.

Tabel IV-4. Data *Tweet* Setelah Dilakukan Proses *Case Folding*

No	<i>Tweet</i>
1	jadwal liga inggris pekan ini akan menyajikan dua derby besar. akan ada derby manchester di old trafford dan derby london utara di markas tottenham hotspur. https://t.co/63qqt8fx0u
2	polres jakbar mengungkap kasus investasi bodong mengatasnamakan waralaba double dipps. total kerugian korban mencapai rp 19,6 miliar. https://t.co/cx7zhorvx1
3	tiket pesawat harganya diklaim sudah turun saat ini. benarkah demikian? https://t.co/vim8pmasnv
4	setelah sebelumnya ada pria jepang yang ingin hidup sebagai hewan, kini ada lagi yang punya keinginan serupa. bedanya, dia mau jadi serigala.

	https://t.co/hry3diecie
5	belakangan ini kasus keracunan nitrogen cair dalam jajanan chiki ngebul tengah ramai diperbincangkan masyarakat. apakah berobatnya bisa menggunakan bpjs? ini jawabannya! https://t.co/enmkuvqxzj

2. *Cleaning*

Proses *cleaning* bertujuan untuk membersihkan data komentar dari karakter seperti simbol, emoticon, dan angka. Hasil *cleaning* dapat dilihat pada Tabel IV-5.

Tabel IV-5. Data *Tweet* Setelah Dilakukan Proses *Cleaning*

No	<i>Tweet</i>
1	jadwal liga inggris pekan ini akan menyajikan dua derby besar akan ada derby manchester di old trafford dan derby london utara di markas tottenham hotspur
2	polres jakbar mengungkap kasus investasi bodong mengatasnamakan waralaba double dipps total kerugian korban mencapai rp miliar
3	tiket pesawat harganya diklaim sudah turun saat ini benarkah demikian
4	setelah sebelumnya ada pria jepang yang ingin hidup sebagai hewan kini ada lagi yang punya keinginan serupa bedanya dia mau jadi serigala
5	belakangan ini kasus keracunan nitrogen cair dalam jajanan chiki ngebul tengah ramai diperbincangkan masyarakat apakah berobatnya bisa menggunakan bpjs ini jawabannya

4.2.3.3 Analisis Proses Pemodelan Topik

Dalam melakukan pemodelan topik, metode *BERTopic* memiliki tiga tahapan yaitu melakukan *document embedding*, *document clustering* untuk melakukan *cluster* ke dalam bentuk *semantic similar cluster*, lalu *topic representation* untuk membuat representasi topik dari masing-masing *cluster*.

Proses *document embedding* dilakukan untuk merepresentasikan suatu kata atau kalimat ke dalam bentuk *dense vector*. Jika suatu dokumen memiliki semantik yang sama dengan dokumen lainnya, maka dapat diasumsikan bahwa dokumen tersebut memiliki topik yang sama. Pada penelitian ini, proses *document embedding* akan dilakukakan menggunakan SBERT. Adapun input data pada proses *document embedding* ini yaitu dataset hasil pra-pengolahan. Setelah dilakukan proses *document embedding*, maka output data yang dihasilkan berupa data *vector embedding*.

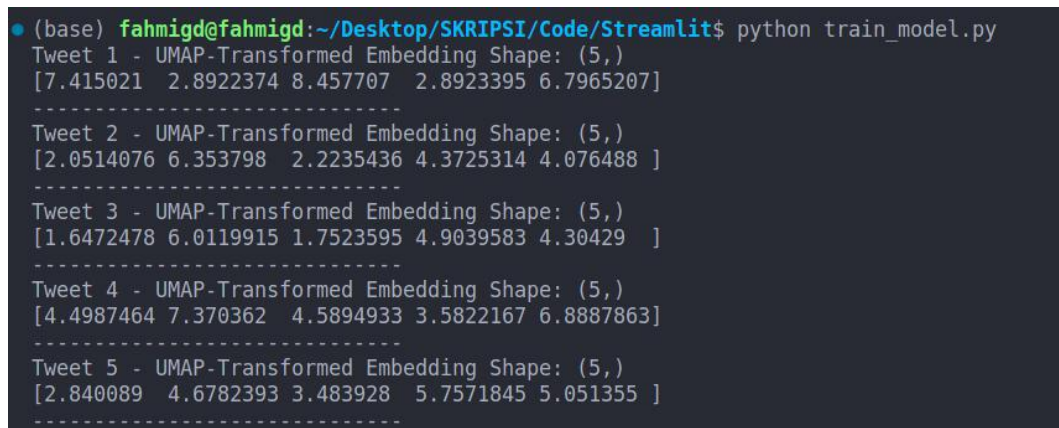
```

(base) fahmigd@fahmigd:~/Desktop/SKRIPSI/Code/Streamlit$ python train_model.py
Tweet 1 - Embedding Shape: (384,)
[ 6.45241663e-02 -3.76279086e-01 1.78568810e-01 8.70958790e-02
 7.08903745e-02 -9.75040530e-05 -5.33179283e-01 2.18153223e-01
 6.55545518e-02 4.83219773e-02 1.21230841e-01 -2.89306920e-02
 1.78266183e-01 -3.72342914e-02 -1.94377989e-01 2.53945868e-02
 1.19039774e-01 -3.31432253e-01 1.84385151e-01 1.58014730e-01
 1.02774529e-02 3.92105728e-02 -3.60586792e-02 4.75431569e-02
 2.30812550e-01 1.83675945e-01 5.14483824e-02 -2.77794059e-02
 1.18623103e-03 6.03749603e-02 -5.11775129e-02 -2.40719505e-02
 -1.71276350e-02 1.21387072e-01 2.15359688e-01 4.23827916e-02
 2.57252008e-01 1.17291115e-01 1.40890375e-01 4.11344040e-03
 -3.17761563e-02 2.77233511e-01 2.69340247e-01 -3.06012064e-01
 1.81055963e-01 -1.27777085e-01 1.87240522e-02 -1.49938539e-01
 -5.46735749e-02 4.48800661e-02 9.97389555e-02 -1.21383674e-01
 -2.39200965e-02 2.07187757e-01 -1.39402062e-01 7.40559995e-02
 1.50678739e-01 -6.18905537e-02 -1.54426843e-01 1.37508333e-01
 5.16690016e-01 2.34240860e-01 3.73167433e-02 2.43803021e-02
 2.26788297e-02 8.00816044e-02 3.46009225e-01 3.72153036e-02
 -8.20953920e-02 6.14079498e-02 8.63246396e-02 -1.01580704e-02
 2.00537249e-01 -2.87821181e-02 -1.94715455e-01 -8.12204406e-02
 -1.43188059e-01 3.22904319e-01 -3.23836625e-01 -1.51531577e-01
 -1.97201828e-03 1.49081245e-01 7.62327760e-03 -6.73412606e-02
 -3.75775248e-02 2.23203242e-01 -2.10529551e-01 -1.47855982e-01
 9.94200706e-02 3.3223834e-03 2.83023238e-01 -9.49484035e-02
 2.43909797e-03 8.41044709e-02 5.86532466e-02 -2.36832038e-01
 -3.70464623e-01 -2.35637948e-01 2.17614904e-01 1.87238343e-02
 -3.42332162e-02 9.50453058e-02 -5.37200691e-03 4.01522249e-01
 2.86170393e-01 3.46018702e-01 9.21033695e-02 6.89565988e-03
 -2.08465964e-01 3.30863744e-01 4.65753190e-02 -3.45800966e-01
 1.56657279e-01 -5.63480100e-03 1.72224596e-01 -2.31829688e-01
 3.31791997e-01 -1.49191739e-02 -4.07429248e-01 3.69036376e-01
 -4.53288713e-03 -4.39481854e-01 -2.67855853e-01 -3.29044610e-02
 3.00423354e-01 -7.12775961e-02 -1.78367659e-01 -6.26631528e-02
 5.67239761e-01 -3.80006284e-01 -1.82739869e-01 1.04871340e-01
 1.34599313e-01 1.19851872e-01 6.75445795e-02 -5.04417680e-02
 1.88179575e-02 9.50003564e-02 -1.30193204e-01 -5.63985333e-03
 -7.27125183e-02 1.14065379e-01 1.65209989e-03 -4.18696273e-03
 -6.00693338e-02 -1.12164252e-01 -4.39005822e-01 1.56743780e-01
 -3.71692739e-02 -3.80971640e-01 5.24541400e-02 -3.00170690e-01
 -1.20565854e-02 1.15569465e-01 -6.44765273e-02 8.02754238e-02
 6.27993867e-02 -3.33770439e-02 -1.27190739e-01 6.91660419e-02

```

Gambar IV-1. Output Proses Document Embedding Menggunakan SBERT

Hasil dari *document embedding* akan meningkatkan dimensi data, sehingga perlu dilakukan *dimensionality reduction*. Pada penelitian ini, proses *dimensionality reduction* akan dilakukan menggunakan metode UMAP. Adapun input pada proses *dimensionality reduction* yaitu *vector embedding* yang berupa *numerical* data hasil dari proses *document embedding* sebelumnya. Setelah dilakukan proses *dimensionality reduction* menggunakan metode UMAP, maka output data yang dihasilkan berupa data *vector embedding* yang dimensi datanya sudah tereduksi.



```

(base) fahmigd@fahmigd:~/Desktop/SKRIPSI/Code/Streamlit$ python train_model.py
Tweet 1 - UMAP-Transformed Embedding Shape: (5,)
[7.415021 2.8922374 8.457707 2.8923395 6.7965207]
-----
Tweet 2 - UMAP-Transformed Embedding Shape: (5,)
[2.0514076 6.353798 2.2235436 4.3725314 4.076488 ]
-----
Tweet 3 - UMAP-Transformed Embedding Shape: (5,)
[1.6472478 6.0119915 1.7523595 4.9039583 4.30429 ]
-----
Tweet 4 - UMAP-Transformed Embedding Shape: (5,)
[4.4987464 7.370362 4.5894933 3.5822167 6.8887863]
-----
Tweet 5 - UMAP-Transformed Embedding Shape: (5,)
[2.840089 4.6782393 3.483928 5.7571845 5.051355 ]
-----

```

Gambar IV-2. *Output Proses Dimensionality Reduction Menggunakan UMAP*

Proses *document clustering* pada penelitian ini dilakukan dengan metode *Hierarchical Density Based Spatial Clustering of Applications with Noise* (HDBSCAN). Metode HDBSCAN sendiri menggunakan pendekatan *soft clustering*, dimana *noise* dimodelkan sebagai *outlier* sehingga dokumen yang tidak terkait tidak akan dimasukkan ke dalam *cluster*. Hal ini tentunya akan meningkatkan representasi topik yang dihasilkan nantinya. Adapun input data pada proses clustering menggunakan metode HDBSCAN ini yaitu *vector embedding* yang dimensinya sudah tereduksi menggunakan metode UMAP. Setelah dilakukan proses *clustering* menggunakan metode HDBSCAN, maka

output data yang dihasilkan berupa *clustering* label dari setiap data teks ulasan pada dataset.

```

• (base) fahmigd@fahmigd:~/Desktop/SKRIPSI/Code/Streamlit$ python train_model.py
Tweet 1 - Cluster Label: 6
Tweet 2 - Cluster Label: -1
Tweet 3 - Cluster Label: 23
Tweet 4 - Cluster Label: 113
Tweet 5 - Cluster Label: 11
-----

```

Gambar IV-3. *Output Proses Document Clustering Menggunakan HDBSCAN*

Dari hasil *cluster* yang diperoleh, setiap *cluster* akan direpresentasikan oleh satu topik. TF-IDF berbasis kelas digunakan untuk membuat merepresentasikan topik pada setiap *cluster*. Penggunaan *c- TF-IDF* akan menghasilkan distribusi topik kata untuk setiap *cluster* dokumen karena metode ini memodelkan pentingnya kata dalam *cluster* dibandingkan dengan dokumen individual. Adapun input data pada proses representasi *cluster* menggunakan *c-TF-IDF* ini yaitu *cluster* data dari proses HDBSCAN sebelumnya. Setelah dilakukan proses representasi *cluster* menggunakan *c-TF-IDF*, maka *output* data yang dihasilkan berupa kata-kata yang merepresentasikan setiap topik atau *cluster*.

```

Topik -1
[('di', 0.007100022059012341), ('yang', 0.007063342370298324), ('via', 0.006236764587103442), ('dan', 0.006205747925817646), ('ini', 0.005863435821344234), ('itu', 0.005748406271818661), ('tahun', 0.005579603206596981), ('dari', 0.00552376985960003), ('saat', 0.005426110405848091), ('untuk', 0.005345523269089128)]
Topik 0
[('aff', 0.055853337610592566), ('indonesia', 0.048608090351541906), ('timnas', 0.039377298986947854), ('thailand', 0.035435443912713686), ('kamboja', 0.028186035211227945), ('laga', 0.027158738665669136), ('garuda', 0.025681989832835754), ('brunei', 0.02473225918221752), ('piala', 0.024366463099072973), ('vs', 0.022483305180813554)]
Topik 1
[('islam', 0.03986337209947637), ('allah', 0.033213427719561706), ('nabi', 0.02935904227901554), ('swt', 0.028665155550490387), ('muslim', 0.028031488293073628), ('al', 0.027480437268809766), ('saw', 0.02654391705525721), ('salat', 0.024935256318817163), ('rasulullah', 0.023286693702677337), ('doa', 0.017930988575983037)]
Topik 2
[('pemilu', 0.06796218573816548), ('partai', 0.0657977456357781), ('kpu', 0.042136652721838155), ('ummat', 0.039017550284369464), ('anies', 0.034872839382557065), ('politik', 0.02976463798454572), ('peserta', 0.02792058247426628), ('baswedan', 0.027600144816221753), ('bawaslu', 0.02479351230997223), ('verifikasi', 0.024751339744318087)]
Topik 3
[('argentina', 0.08842506518870799), ('final', 0.04636709680721405), ('piala', 0.04492276663361127), ('dunia', 0.04253811495138676), ('prancis', 0.04134419354860461), ('juara', 0.033179710637643224), ('kemenangan', 0.029091650866068495), ('tango', 0.02847058146668498), ('tim', 0.026037025948231672), ('messi', 0.0251168588640949)]
Topik 4
[('makanan', 0.03830988474242838), ('restoran', 0.038165483241337256), ('kopi', 0.031000326251771376), ('makan', 0.025854214887326497), ('diner', 0.01796663440865707), ('karen', 0.01772436937230359), ('menu', 0.017080010956352018), ('nasi', 0.016151010864221757), ('ini', 0.01510242327534907), ('kafe', 0.014717426660098575)]

```

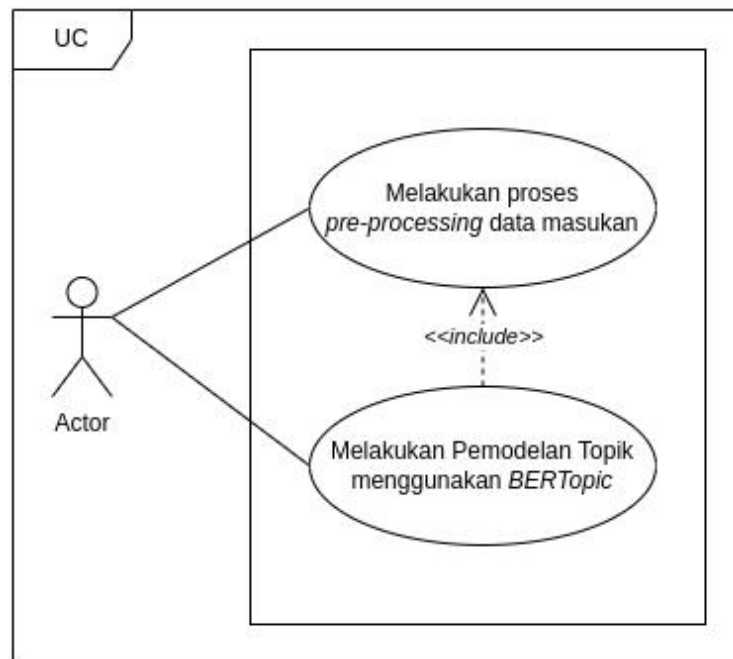
Gambar IV-4. *Output Proses Topic Representation Menggunakan c-TF-IDF*

4.2.3.4 Analisis Hasil Pemodelan Topik

Pada pengembangan perangkat lunak, hasil pemodelan topik yang diperoleh dianalisis menggunakan *coherence score cv*. Tujuan analisis hasil pemodelan topik adalah untuk mengetahui kualitas topik dari hasil pemodelan topik. Semakin tinggi *coherence score cv* menunjukkan semakin informatif atau bermakna topik tersebut. Dan sebaliknya semakin rendah *coherence score cv* maka semakin tidak informatif atau tidak bermakna topik tersebut.

4.2.4 Implementasi

Berdasarkan kebutuhan dan analisis yang telah dilakukan pada fase sebelumnya, desain perangkat lunak akan dimodelkan menggunakan diagram *use case*. Diagram *use case* perangkat lunak digambarkan pada Gambar IV-5



Gambar IV-5. *Use Case* Pemodelan Topik Menggunakan *BERTopic*

Penjelasan rinci diagram *use case* direpresntasikan pada tabel definisi aktor dan definisi *use case* dimuat pada tabel IV- 6 dan IV-7.

Tabel IV-6. Tabel Definisi Aktor

No	Aktor	Definisi
1	Pengguna	Seseorang berinteraksi dan mengoperasikan perangkat lunak

Tabel IV-7. Definisi Use Case

No	Aktor	Definisi
1	Melakukan Proses <i>Pre-processing</i> Data Masukan	Proses ini bertujuan untuk melakukan proses pra-pengolahan data masukan
2	Melakukan Pemodelan Topik Menggunakan <i>BERTopic</i>	Proses ini bertujuan untuk melakukan pemodelan topik pada data masukan

Berdasarkan tabel di atas, deskripsi skenario *use case* ditunjukkan pada Tabel IV-8, Tabel IV-9.

Tabel IV-8. Skenario *Use Case* Melakukan Proses *Pre-processing* Data Masukan

Identifikasi	
Nomor Use Case	01
Nama Butir Uji	Melakukan Proses <i>Pre-processing</i> Data Masukan
Aktor	User
Tujuan	Melakukan pra-pengolahan data masukan
Deskripsi	Proses ini bertujuan untuk melakukan pra-pengolahan data masukan
Kondisi Awal	Belum ada data yang dimasukkan

Skenario Normal	
Aktor	Sistem
1. User menekan tombol “Load File .csv”	
	2. Menampilkan jendela pencarian berkas
3. Memilih file dataset yang akan diproses	
	4. Melakukan proses pra-pengolahan dataset yang dimasukkan user
	5. Menampilkan data hasil pra-pengolahan dan sebelum pra-pengolahan
Kondisi Akhir Skenario Normal : Menampilkan Data Sebelum Pra-pengolahan dan Data Setelah Pra-pengolahan	

Tabel IV-9. Skenario *Use Case* Melakukan Pemodelan Topik Menggunakan *BERTopic*

Identifikasi	
Nomor Use Case	02
Nama Butir Uji	Melakukan Pemodelan Topik Menggunakan <i>BERTopic</i>
Aktor	User
Tujuan	Melakukan pemodelan topik pada data masukan
Deskripsi	Proses ini bertujuan untuk melakukan pemodelan topik pada data masukan

Kondisi Awal	Proses pra-pengolahan data selesai
Skenario Normal	
Aktor	Sistem
1. User menekan tombol “Extract Topic”	
	2. Melakukan pemodelan topik pada data hasil pra-pengolahan
	3. Menampilkan hasil pemodelan topik
Kondisi Akhir Skenario Normal : Menampilkan Hasil Pemodelan Topik	

4.3 Fase Elaborasi

Fase kedua pada proses pengembangan adalah identifikasi sistem. Proses yang dilakukan antara lain perancangan data dan tampilan serta pemodelan diagram aktivitas dan diagram alur.

4.3.1 Pemodelan Bisnis

Pada subbab ini akan membahas mengenai perancangan perangkat lunak. Perancangan yang dilakukan antara lain perancangan data dan perancangan tampilan pengguna yang akan digunakan.

4.3.1.1 Perancangan Data

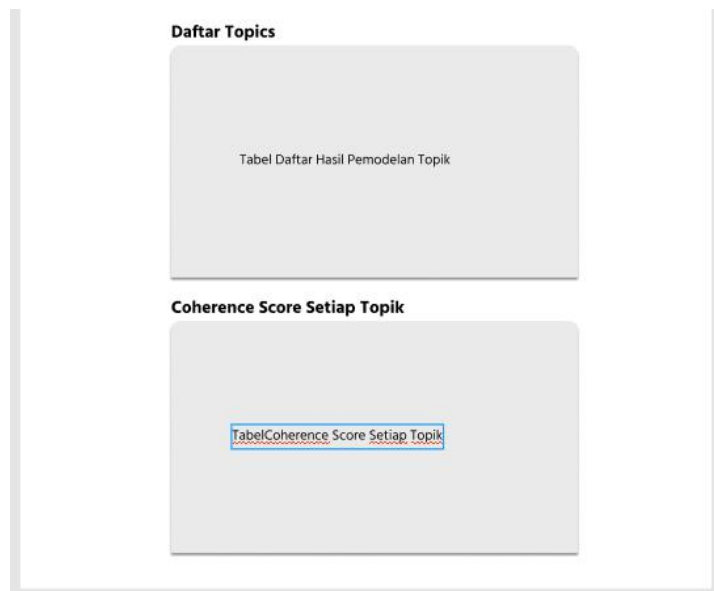
Perangkat lunak yang dikembangkan bertujuan untuk mengekstraksi topik topik yang ada pada sekumpulan data *tweet* bahasa Indonesia. Data masukan yang digunakan berupa sebuah dataset *tweet* berformat .csv.

4.3.1.2 Perancangan Antarmuka

Perancangan antarmuka atau tampilan bertujuan untuk merancang tampilan perangkat lunak sehingga perangkat lunak yang dikembangkan sesuai dengan kebutuhan pengguna. Adapun rancangan tampilan perangkat lunak digambarkan pada Gambar IV-6 dan Gambar IV-7



Gambar IV-6. Rancangan Antarmuka Pra-Pengolahan Data



Gambar IV-7. Rancangan Antarmuka Hasil Pemodelan Topik

4.3.2 Kebutuhan

Pada subbab ini akan menjelaskan kebutuhan sistem dalam proses pengembangan dan pembangunan perangkat lunak. Pengembangan perangkat lunak membutuhkan perangkat lunak (*software*), perangkat keras (*hardware*), dan bahasa pemrograman. Bahasa pemrograman Python digunakan dalam pengembangan perangkat lunak pada penelitian ini.

Perangkat keras yang dibutuhkan adalah sebagai berikut :

1. Processor : AMD® Ryzen 5 3550h with radeon vega mobile gfx × 8
2. RAM : 8 GB
3. SSD : 512 GB

Sedangkan perangkat lunak yang dibutuhkan adalah :

1. Sistem Operasi : Ubuntu 20.04.6 LTS
2. Teks Editor : Visual Studio Code, Google Colab
3. Browser : Chrome

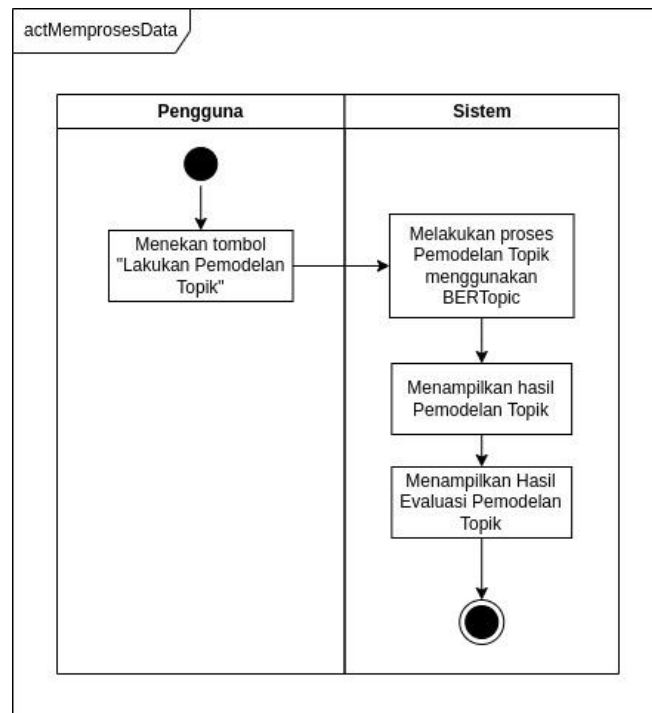
4.3.3 Analisis dan Perancangan

Pada subbab ini akan menjelaskan perancangan yang perlu dilakukan dalam pengembangan perangkat lunak. Rancangan tersebut mencakup rancangan diagram aktivitas dan diagram alur.

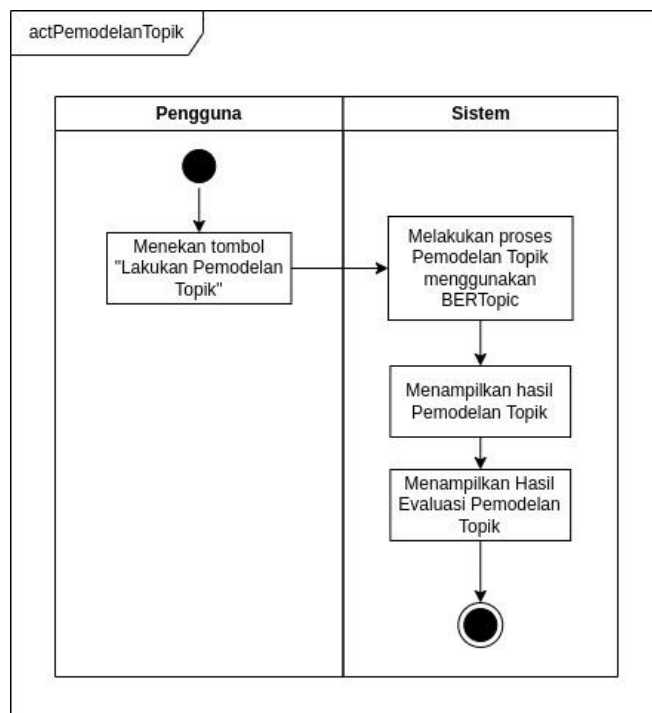
4.3.3.1 Diagram Aktivitas

Diagram aktivitas menggambarkan aliran aktivitas pada perangkat lunak. Berdasarkan use case yang telah dibuat, terdapat dua diagram aktivitas pada perangkat lunak. Diagram aktivitas pada Gambar IV-8 menunjukkan aktivitas

pengguna dalam memproses data masukan pada sistem. Diagram aktivitas proses Pemodelan Topik menggunakan *BERTopic* ditunjukkan pada Gambar IV-9.



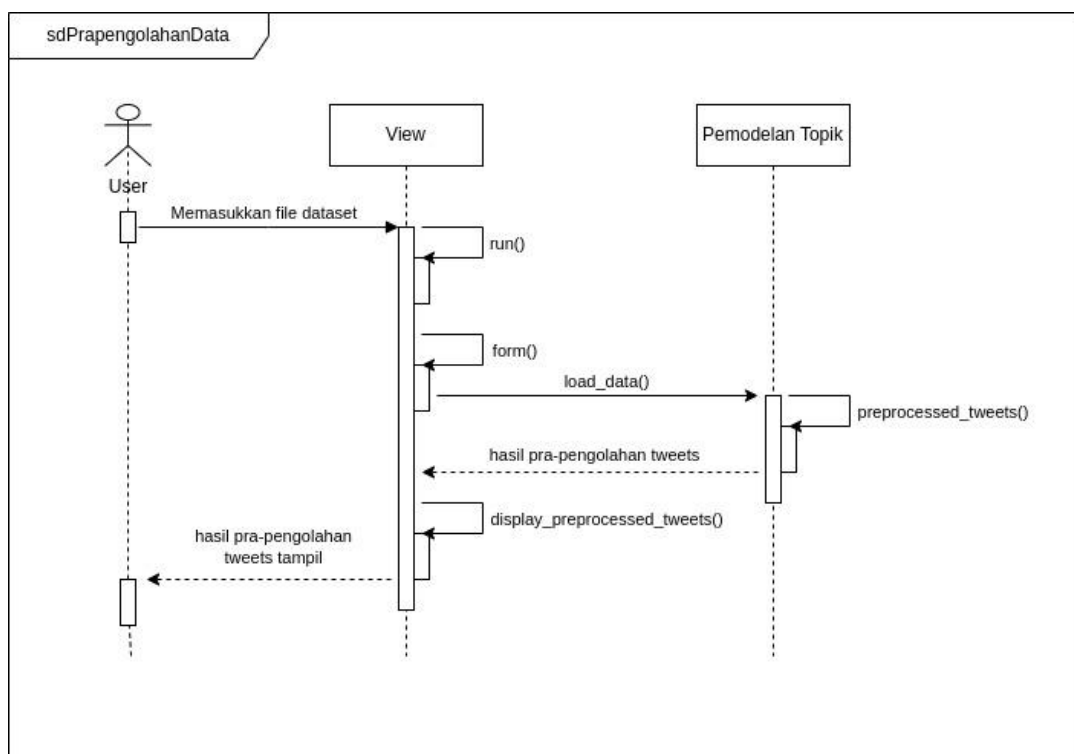
Gambar IV-8. Diagram Aktivitas Melakukan Pra-Pengolahan Data Pada Sistem



Gambar IV-9. Diagram Aktivitas Melakukan Pemodelan Topik

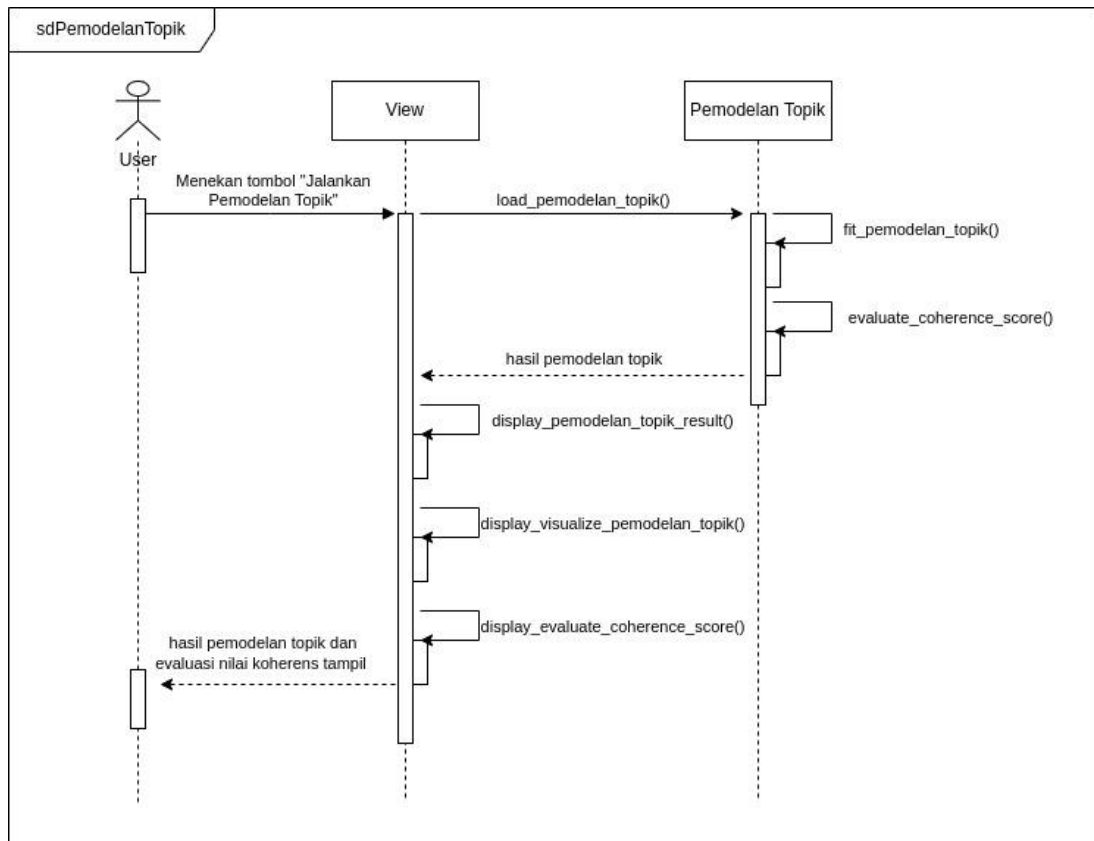
4.3.3.2 Diagram Alur

Diagram alur menggambarkan alur atau urutan interaksi antar objek pada perangkat lunak berdasarkan waktu. Berdasarkan *use case* yang telah dibuat terdapat dua diagram alur perangkat lunak yang ditunjukkan pada Gambar IV-10 dan Gambar IV-11.



Gambar IV-10. Diagram Alur Proses Pra-Pengolahan Data

Gambar IV-10 diatas merupakan diagram alur yang menjelaskan semua alur dari pra-pengolahan data. Alur tersebut diawali dengan pengguna menekan tombol *Browse files* yang nantinya pengguna akan melihat *window* pencarian data dan langsung memasukkan file dataset yang akan diproses. Keluaran dari proses ini adalah sebuah list yang ditampilkan dalam sebuah tabel.



Gambar IV-11. Diagram Alur Proses Pemodelan Topik Menggunakan BERTopic

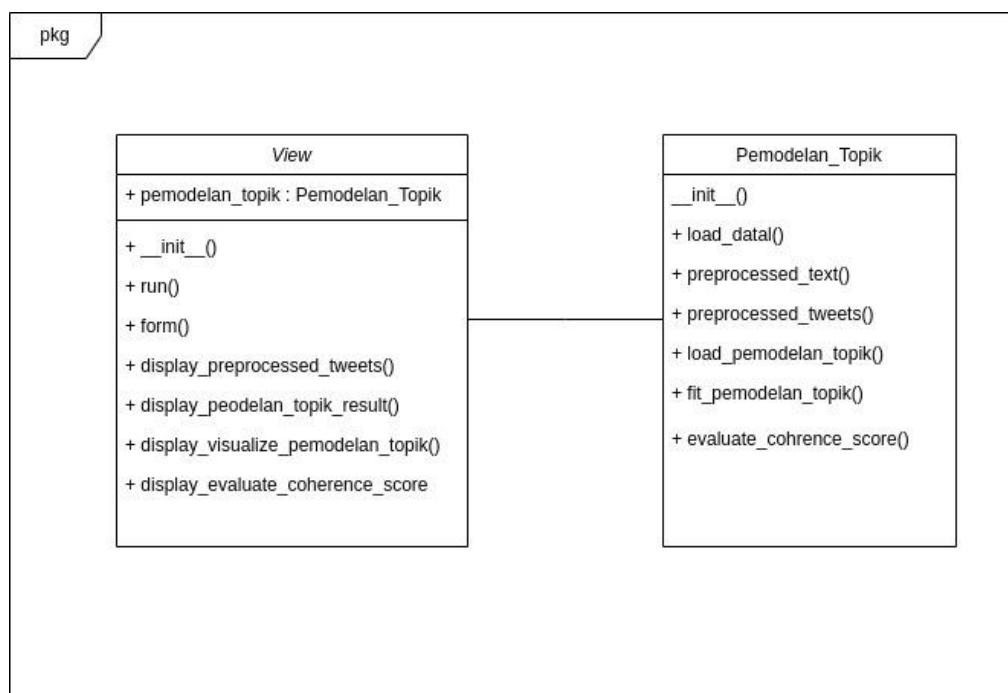
Gambar IV-11 menunjukkan diagram alur yang menjelaskan urutan interaksi pada proses Pemodelan Topik menggunakan BERTopic. Diawali dengan pengguna menekan tombol “Jalankan Pemodelan Topik”. Kemudian perangkat lunak akan menjalankan proses pemodelan topik pada dataset hasil pra-pengolahan menggunakan model *BERTopic*. Hasil proses ini berupa daftar topik dan nilai koherensi setiap topik.

4.4 Fase Konstruksi

Subbab ini akan menjelaskan bagian inti serta fitur lain dari perangkat lunak yang dikembangkan. Hasil dari fase ini berupa sebuah perangkat lunak yang akan dijadikan alat pada penelitian.

4.4.1 Kebutuhan

Subbab ini menjelaskan pemodelan perangkat lunak dalam bentuk diagram kelas. Diagram kelas digunakan untuk menggambarkan kelas yang dirancang dan dibuat dalam pengembangan perangkat lunak serta relasi antara masing-masing kelas sehingga menjadi sebuah sistem. Kelas yang dibentuk meliputi 2 kelas, 1 kelas tampilan dalam 1 file View.py (View class), 1 kelas pemodelan topik dalam file Pemodelan_Topik.py (Pemodelan_Topik class). Gambar IV-12 dibawah ini akan merepresentasikan hubungan dari kelas-kelas kedalam bentuk diagram kelas perangkat lunak.



Gambar IV-12. Diagram Kelas Perangkat Lunak

4.4.2 Implementasi

Pada fase konstruksi ini rancangan yang telah dibuat akan diterapkan menjadi sebuah perangkat lunak. Proses pengembangan perangkat lunak dalam

penelitian ini menggunakan beberapa pustaka yang terdapat pada bahasa Python antara lain Stremalit, Gensim dan Pandas.

4.4.2.1 Implementasi Kelas

Kelas-kelas yang telah dirancang diterapkan menggunakan bahasa pemrograman Python. Penerapan kelas dalam bahasa Python ditunjukkan pada Tabel IV-13.

Tabel IV-13. Keterangan Implementasi Kelas

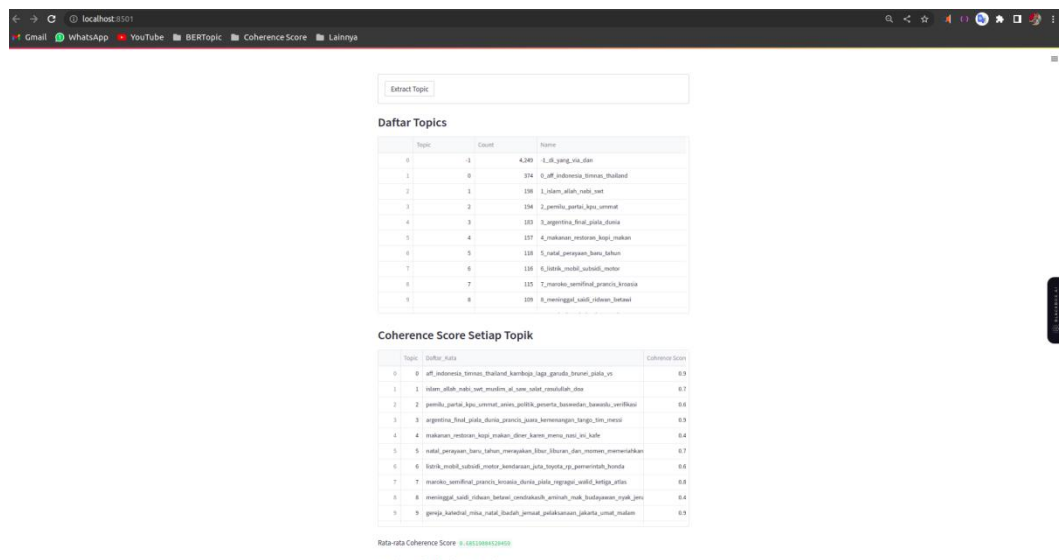
No	Nama Kelas	Nama File	Keterangan
1	View	View.py	Kelas ini menangani pembuatan dan pengaturan objek tampilan pada perangkat lunak
2	Pemodelan_Topik	Pemodelan_Topik.py	Kelas ini merupakan kelas yang menangani proses-proses pada sistem

4.4.2.2 Implementasi Antarmuka

Subbab ini menunjukkan implementasi antarmuka perangkat lunak. Implementasi dilakukan berdasarkan rancangan antarmuka pada fase elaborasi. Tampilan perangkat lunak pengujian ditunjukkan pada Gambar IV-14 dan Gambar IV-15.



Gambar IV-14. Implementasi Antarmuka Pra-Pengolahan Data



Gambar IV-15. Implementasi Antarmuka Hasil Pemodelan Topik

4.5 Fase Transisi

Fase ini berisi pengujian perangkat lunak berdasarkan hasil pengembangan perangkat lunak pada fase konstruksi.

4.5.1 Pemodelan Bisnis

Pengujian *blackbox* digunakan pada pengujian perangkat lunak. Rencana pengujian dibuat berdasarkan *use case* yang dirancang pada tahap insepisi.

4.5.2 Kebutuhan

Alat yang digunakan dalam pengujian perangkat lunak sama dengan alat ada pembuatan perangkat lunak. Perangkat keras yang digunakan adalah sebagai berikut:

1. Processor : AMD® Ryzen 5 3550h with radeon vega mobile gfx × 8
2. RAM : 8 GB
3. SSD : 512 GB

Sedangkan perangkat lunak yang dibutuhkan adalah :

1. Sistem Operasi : Ubuntu 20.04.6 LTS
2. Teks Editor : Visual Studio Code, Google Colab
3. Browser : Chrome

4.5.3 Analisis dan Perancangan

Subbab ini akan membahas tentang rencana pengujian perangkat lunak berdasarkan use case yang telah dibuat.

4.5.3.1 Rencana Pengujian

Rencana pengujian pemodelan topik pada tweet bahasa Indonesia ditunjukkan pada beberapa tabel di bawah ini. Skenario pengujian *blackbox* perangkat lunak dapat dilihat pada Tabel IV-11 dan Tabel IV-12

1. Rencana Pengujian *Use Case* Proses Pra-Pengolahan Data

Tabel IV-11. Rencana Pengujian *Use Case* Proses Pra-Pengolahan Data

No	Identifikasi	Pengujian	Tingkat Pengujian
1	U-1	Memuat dan melakukan proses pra-pengolahan pada data masukan pengguna	Pengujian Unit

2. Rencana Pengujian *Use Case* Proses Pemodelan Topik Menggunakan *BERTopic*

Tabel IV-12. Rencana Pengujian *Use Case* Proses Pemodelan Topik Menggunakan *BERTopic*

No	Identifikasi	Pengujian	Tingkat Pengujian
1	U-2	Melakukan pemodelan topik pada dataset tweet menggunakan <i>BERTopic</i>	Pengujian Unit

Berdasarkan dua rencana pengujian di atas akan diimplementasikan pengujian menggunakan perangkat lunak yang telah dikembangkan. Penjelasan detail tentang rencana pengujian diuraikan dengan variabel antara lain identifikasi, deksripsi, prosedur pengujian, masukan, keluaran yang diharapkan, kriteria evaluasi hasil, hasil yang diperoleh, dan diakhiri dengan kesimpulan apakah perangkat lunak tersebut memenuhi kriteria pengujian atau tidak.

4.5.3.2 Implementasi

Tahapan ini akan menjelaskan uji kasus berdasarkan rencana pengujian sebelumnya.

1. Pengujian *Use Case* Proses Pra-Pengolahan Data

Tabel IV-13. Pengujian *Use Case* Proses Pra-Pengolahan Data

Identifikasi	Deskripsi	Prosedur Pengujian	Masukan	Keluaran yang Diharapkan	Kriteria Evaluasi Hasil	Hasil yang Diperoleh	Kesimpulan
U-1	Melakukan Proses Pra-Pengolahan Data	Menekan Tombol “Browse files”	Data	Hasil proses pra-pengolahan pada tabel	Perangkat lunak memuat dan melakukan pra-pengolahan data, kemudian menampilkan hasil pra-pengolahan pada tabel di perangkat lunak	Perangkat lunak memuat dan melakukan pra-pengolahan data, kemudian menampilkan hasil pra-pengolahan pada tabel di perangkat lunak	Terpenuhi

2. Rencana Pengujian *Use Case* Proses Pemodelan Topik Menggunakan *BERTopic*

Tabel IV-14. Pengujian *Use Case* Proses Pemodelan Topik Menggunakan *BERTopic*

Identifikasi	Deskripsi	Prosedur Pengujian	Masukan	Keluaran yang Diharapkan	Kriteria Evaluasi Hasil	Hasil yang Diperoleh	Kesimpulan
U-2	Melakukan Proses Pemodelan Topik Menggunakan <i>BERTopic</i>	Menekan Tombol “Lakukan Pemodelan Topik”	Data Hasil Pra-Pengolahan	Hasil Pemodelan Topik	Perangkat lunak melakukan pemodelan topik, kemudian menampilkan hasil pemodelan topik	Perangkat lunak melakukan pemodelan topik, kemudian menampilkan hasil pemodelan topik	Terpenuhi

4.6 Kesimpulan

Bab ini telah menjelaskan secara rinci proses pengembangan perangkat lunak sebagai alat penunjang dalam penelitian pemodelan topik pada *tweet* bahasa Indonesia menggunakan *BERTopic*. Alur pengembangan perangkat lunak pada penelitian ini telah diuraikan sehingga dapat menghasilkan perangkat lunak sesuai dengan kebutuhan penelitian.

BAB V

HASIL DAN ANALISIS

5.1 Pendahuluan

Pada bab ini akan dibahas hasil penelitian menggunakan perangkat lunak yang telah dikembangkan pada bab sebelumnya. Hasil penelitian akan disajikan dengan menggunakan Tabel III-3 dan akan diberikan juga analisis dari hasil penelitian yang telah dilakukan..

5.2 Hasil Penelitian

Penelitian yang telah dilakukan bertujuan untuk melakukan pemodelan topik pada tweet bahasa Indonesia. Hasil pemodelan topik disajikan pada Tabel V-1.

Tabel V-1. Hasil Pemodelan Topik Menggunakan *BERTopic*

Topik	Jumlah Tweet	Daftar Kata
-1	4249	di_yang_via_dan
0	374	aff_indonesia_timnas_thailand
1	198	islam_allah_nabi_swt
2	194	pemilu_partai_kpu_ummat
3	183	argentina_final_piala_dunia
4	157	makanan_restoran_kopi_makan
....

114	16	bunga_acuan_bank_kredit
115	16	baim_wong_prank_kdrt
116	16	jubah_qatar_bisht_lionel
117	16	kanjuruhan_tragedi_mahfud_ham
118	16	taksi_penumpang_halim_bandara

Setelah memperoleh hasil pemodelan topik, topik-topik tersebut di evaluasi menggunakan *coherence score cv*. Tujuan evaluasi ini adalah untuk mengetahui kualitas topik dari hasil pemodelan topik. Semakin tinggi *coherence score cv* nya maka semakin bermakna atau informatif topik tersebut. Sebaliknya semakin rendah *coherence score cv* suatu topik maka semakin tidak bermakna atau tidak informatif topik tersebut. Hasil evaluasi pemodelan topik ditunjukkan pada Tabel V-2.

Tabel V-2. Hasil Evaluasi Pemodelan Topik Menggunakan *BERTopic*

Topik	Daftar Kata	<i>Coherence Score cv</i>
47	ular_alprie_panji_petualang_priyono_kobra_king_cobra_asisten_dipatuk	0.995
57	umbaran iptu wartawan kapolsek tvri intel krade nan menyamar wibowo blora	0.994
101	toleransi_sawah_antar_kampung_beragama_kerukunan umat pondasi ditanamkan harmoni	0.990
112	kloset_malang_bilik_mcc_mandi_kamar_creative_center luar nyeleneh	0.990
89	selle_nirwana_morowali_gni_tiktok_seleb_kebakaran_smelter_utara_pt	0.986
....	
25	monyet_zoo_hewan_kandang_mini_mati_bogor_kurus_ragunan_juve	0.279
109	foto_deretan_pedangdut_hantu_lihat_mella_rossa_kulineran_fotografi_jadul	0.263

110	planet_bumi_ilmuwan_bulat_universe_asteroid_ja mes_dceu_astronomi_dihuni	0.215
85	etle_google_mobile_iphone_aplikasi_ponsel_lei_ch atgpt_jun_bca	0.148
107	rsdc_wisma_atlet_kemayoran_majalah_sevilla_mo mbi_isco_mengakhiri_sepakat	0.119

5.3 Analisis Hasil Penelitian

Setelah dilakukan pemodelan topik pada 10.000 *tweet* diperoleh 119 topik utama dan 1 topik *outlier* yang ditunjukkan pada Tabel V-1. Dari 10.000 *tweet* terdapat 5751 *tweet* yang mewaliki topik utama dan 4249 *tweet* yang mewakili topik *outlier*. Besarnya jumlah *tweet* yang mewakili topik *outlier* disebabkan karena dataset *tweet* yang diperoleh dari *tweet* detikcom mencakup berbagai topik yang sangat beragam hal ini mengakibatkan model pemodelan topik mengalami kesulitan untuk menggambarkan topik utama secara rinci, sehingga beberapa kelompok *tweet* dianggap sebagai topik *outlier*.

Berdasarkan hasil evaluasi pemodelan topik yang ditunjukkan pada Tabel V-2, diperoleh rata-rata *coherence score cv* untuk pemodelan topik yaitu 0,685. *Coherence score cv* tertinggi 0,995 untuk topik 47 dan *coherence score cv* terendah 0,119 untuk topik 107. Rendahnya *coherence score cv* pada topik 107 disebabkan karena pada topik tersebut mengandung kata-kata yang tidak relevan satu sama lain. Ketidakrelevanan kata-kata dalam topik dapat menyebabkan rendahnya kohesifitas dan kesulitan bagi model untuk menggambarkan topik secara akurat.

5.4 Kesimpulan

Berdasarkan Hasil dan analisis hasil pemodelan topik pada *tweet* bahasa Indonesia menggunakan BERTopic telah dilakukan pada bab ini. Pada penelitian ini diperoleh 119 topik utama dengan *coherence score cv* tertinggi 0,995 pada topik 47.

BAB VI

KESIMPULAN DAN SARAN

6.1 Pendahuluan

Pada bab ini membahas mengenai kesimpulan dan saran dari uraian bab sebelumnya serta berdasarkan analisis hasil penelitian. Sehingga dapat dijadikan acuan untuk penelitian berikutnya pada bidang ini..

6.2 Kesimpulan

Berdasarkan penelitian yang telah dijelaskan pada bab sebelumnya, diambil kesimpulan sebagai berikut.

1. Pemodelan topik pada *tweet* bahasa Indonesia menggunakan *BERTopic* berhasil dikembangkan dengan baik.
2. Program dapat melakukan pemodelan topik pada 10.000 *tweet* bahasa Indonesia dan mendapatkan 119 topik utama dengan *coherence score cv* tertinggi 0,995 untuk topik 47 dan *coherence score cv* terendah 0,119 untuk topik 107, adapun rata-rata *coherence score cv* yaitu 0,685.

6.3 Saran

Pada penelitian selanjutnya disarankan untuk menggunakan metode pra-pengolahan teks yang berbeda pada dataset yang digunakan, serta mencoba menggunakan metode *document embedding*, *dimensionality reduction* dan *document clustering* yang berbeda.

DAFTAR PUSTAKA

- Abuzayed, A., & Al-Khalifa, H. (2021). BERT for Arabic topic modeling: an experimental study on BERTopic technique. *Procedia Computer Science*, 189, 191-194.
- Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001, January). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory* (pp. 420-434). Springer, Berlin, Heidelberg.
- Allaoui, M., Kherfi, M. L., & Cheriet, A. (2020, June). Considerably improving clustering algorithms using UMAP dimensionality reduction technique: a comparative study. In *International Conference on Image and Signal Processing* (pp. 317-325). Springer, Cham.
- Al-khairi, Y. U., Wibisono, Y., & Putro, B. L. (2017). Deteksi topik fashion pada twitter dengan latent dirichlet allocation.
- Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.
- Arora, S., Ge, R., & Moitra, A. (2012, October). Learning topic models--going beyond SVD. In *2012 IEEE 53rd annual symposium on foundations of computer science* (pp. 1-10). IEEE.
- Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999, January). When is “nearest neighbor” meaningful?. In *International conference on database theory* (pp. 217-235). Springer, Berlin, Heidelberg.

- Blei, D. M. (2013). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.
- Carbonetto, P., Sarkar, A., Wang, Z., & Stephens, M. (2021). Non-negative matrix factorization algorithms greatly improve topic model fits. *arXiv preprint arXiv:2105.13440*.
- Chilmi, M. L. C. (2021). Latent dirichlet allocation lda untuk mengetahui topik pembicaraan warganet twitter tentang omnibus law (Bachelor's thesis, Fakultas Sains Dan Teknologi UIN Syarif Hidayatullah Jakarta).
- Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, 7.
- Gornik, D. (2004). IBM Rational Unified Process: Best practices for software development teams. *Rational Software White Paper TP026B, Rev, 11(01)*.
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- Hendry, D., Darari, F., Nurfadillah, R., Khanna, G., Sun, M., Condylis, P. C., & Taufik, N. (2021, October). Topic modeling for customer service chats. In *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (pp. 1-6). IEEE.
- Joachims, T. (1996). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. *Carnegie-mellon univ pittsburgh pa dept of computer science*.

- Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11), 15169-15211.
- Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788-791.
- McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- McInnes, L., & Healy, J. (2017, November). Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 33-42). IEEE.
- Meeks, E., & Weingart, S. B. (2012). The digital humanities contribution to topic modeling. *Journal of Digital Humanities*, 2(1), 1-6.
- Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010, June). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics* (pp. 100-108).
- Newman, D., Noh, Y., Talley, E., Karimi, S., & Baldwin, T. (2010, June). Evaluating topic models for digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries* (pp. 215-224).
- Pandove, D., Goel, S., & Rani, R. (2018). Systematic review of clustering high-dimensional and large datasets. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(2), 1-68.

- Patmawati, P., & Yusuf, M. (2021). Analisis Topik Modelling Terhadap Penggunaan Sosial Media Twitter oleh Pejabat Negara. *Building of Informatics, Technology and Science (BITS)*, 3(3), 122-129.
- Pradha, S., Halgamuge, M. N., & Vinh, N. T. Q. (2019, October). Effective text data preprocessing technique for sentiment analysis in social media data. In 2019 11th international conference on knowledge and systems engineering (KSE) (pp. 1-8). IEEE.
- Putra, I. M. K. B. (2017). Analisis topik informasi publik media sosial di surabaya menggunakan pemodelan latent dirichlet allocation (LDA) (Doctoral dissertation, Institut Teknologi Sepuluh Nopember).
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Sánchez-Franco, M. J., & Rey-Moreno, M. (2022). Do travelers' reviews depend on the destination? An analysis in coastal and urban peer-to-peer lodgings. *Psychology & Marketing*, 39(2), 441-459.
- Steinbach, M., Ertöz, L., & Kumar, V. (2004). The challenges of clustering high dimensional data. In *New directions in statistical physics* (pp. 273-309). Springer, Berlin, Heidelberg.
- Thakur, N., Reimers, N., Daxenberger, J., & Gurevych, I. (2020). Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *arXiv preprint arXiv:2010.08240*.
- Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, 94, 101582.

LAMPIRAN

1. Kode Program dan Dataset

https://github.com/fahmigd/Skripsi_Pemodelan_Topik