**Pemodelan Topik Pemilu 2024 Menggunakan Metode BERTopic pada Artikel Berita Detik.com**

### **SKRIPSI**

#### Dini Aryani

#### 20200040030



**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK, KOMPUTER DAN DESAIN**

**SUKABUMI**

**MEI 2024**

**Pemodelan Topik Pemilu 2024 Menggunakan Metode BERTopic pada Artikel Berita Detik.com**

### **SKRIPSI**

*Diajukan Untuk Memenuhi Salah Satu Syarat Dalam Menempuh*

*Gelar Sarjana Komputer*

#### Dini Aryani

##### 20200040030



**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK, KOMPUTER DAN DESAIN**

**SUKABUMI**

**MEI 2024**

### PERNYATAAN PENULIS

JUDUL = Pemodelan Topik Pemilu 2024 Menggunakan Metode BERTopic pada Artikel Berita Detik.com

NAMA = Dini Aryani

NIM = 20200040030

“Saya menyatakan dan bertanggungjawab dengan sebenarnya bahwa Skripsi ini adalah hasil karya saya sendiri kecuali cuplikan dan ringkasan yang masing-masing telah saya jelaskan sumbernya. Jika pada waktu selanjutnya ada pihak lain yang mengklaim bahwa Skripsi ini sebagai karyanya, yang disertai dengan bukti- bukti yang cukup, maka saya bersedia untuk dibatalkan gelar Sarjana Komputer saya beserta segala hak dan kewajiban yang melekat pada gelar tersebut”

Sukabumi, ……….. Mei 2024

Materai

Dini Aryani

Penulis

**PENGESAHAN SKRIPSI**

JUDUL = Pemodelan Topik Pemilu 2024 Menggunakan Metode BERTopic pada Artikel Berita Detik.com

NAMA = Dini Aryani

NIM = 20200040030

Skripsi ini telah diujikan dan dipertahankan di depan Dewan Penguji pada Sidang Skripsi tanggal ....... Mei 2024 Menurut pandangan kami, Skripsi ini memadai dari segi kualitas untuk tujuan penganugerahan gelar Sarjana Komputer.

Sukabumi, ……. Mei 2024

|  |
| --- |
| Pembimbing 2 |
|  |
|  |
| Alun Sujjada, S.Kom., M.T |
| NIDN : |

|  |
| --- |
| Pembimbing 1 |
|  |
|  |
| Ivana Lucia Kharisma, M.Kom |
| NIDN : |

|  |
| --- |
| Ketua Program Studi Teknik Informatika |
|  |
|  |
| Somantri, S.T., M.Kom |
| NIDN : |

|  |
| --- |
| Ketua Penguji |
|  |
|  |
| ………………… |
| NIDN : |

|  |
| --- |
| Dekan Fakultas Teknik Komputer dan Desain |
|  |
|  |
| …………………… |
| NIDN : |

***Skripsi ini kutujukan kepada Ayahanda dan Ibunda tercinta, Kakak dan Adikku tersayang ……………………………………………………………………………………………………..***

*ABSTRAK*

Pada tahun 2024, Indonesia menggelar pemilihan umum serentak yang diwarnai oleh partisipasi dominan dari generasi muda, terutama generasi Z dan milenial. Kedua generasi ini cenderung mencari informasi politik melalui *internet*, menunjukkan peran penting media digital dalam membentuk opini publik terkait pemilu. Penelitian sebelumnya menunjukkan bahwa eksposur terus-menerus terhadap media digital positif dapat membentuk dukungan dan partisipasi politik dari generasi milenial. Ketersediaan media massa *online* memudahkan interaksi dengan berita aktual bagi semua kalangan. Namun, melacak dan menganalisis topik-topik berita menjadi tantangan, terutama dengan volume besar artikel yang dipublikasikan setiap hari. Oleh karena itu, penelitian ini akan menggunakan teknik *Topic* *Modelling* untuk memodelkan topik-topik terkait pemilu 2024 dari artikel berita Kompas.com. Studi ini akan mengimplementasikan metode BERTopic dalam menganalisis meta deskripsi artikel.

Kata Kunci: *Topic* *modelling*, BERTopic, pemilu 2024, artikel berita, Detik.com

*ABSTRACT*

*In 2024, Indonesia held simultaneous general elections marked by the dominant participation of young generations, particularly Generation Z and millennials. These generations tend to seek political information through the internet, indicating the significant role of digital media in shaping public opinion related to elections. Previous research has shown that continuous exposure to positive digital media can shape the political support and participation of millennials. The availability of online mass media facilitates interaction with current news for all groups. However, tracking and analyzing news topics pose a challenge, especially with the large volume of articles published daily. Therefore, this research will use Topic Modelling techniques to model topics related to the 2024 elections from Kompas.com news articles. This study will implement the BERTopic method in analyzing article titles and descriptions.*

*Keywords: Topic modelling, BERTopic, 2024 election, news, Detik.com*

**KATA PENGANTAR**

Puji syukur kami panjatkan ke hadirat Allah SWT, berkat rahmat dan karunia-Nya akhirnya penulis dapat menyelesaikan skripsi …….dan seterusnya. Tujuan penulisan skripsi ini adalah ………………….dan seterusnya.

Sehubungan dengan itu penulis menyampaikan penghargaan dan ucapan terima kasih yang

sebesar-besarnya kepada :

1. Bapak Rektor Universitas Nusa Putra Sukabumi ………. dan seterusnya.
2. Wakil Rektor I Bidang Akademik Universitas Nusa Putra Sukabumi ……. dan seterusnya.
3. Kepala Program Studi ............ Universitas Nusa Putra Sukabumi ……. dan seterusnya.
4. Dosen Pembimbing I Universitas Nusa Putra Sukabumi ………. dan seterusnya.
5. Dosen Pembimbing II Universitas Nusa Putra Sukabumi ………. dan seterusnya.
6. Dosen Penguji .............................., ................................dan setrusnya.
7. Para Dosen Program Studi ..... Universitas Nusa Putra Sukabumi ………. dan seterusnya.
8. Orang tua dan keluarga .................... dan seterusnya.
9. Rekan –rekan mahasiswa ..................... dan seterusnya.
10. Pihak terkait .......................dan seterusnya.

Penulis menyadari bahwa skripisi ini masih jauh dari kesempurnaan, oleh karena itu kritik dan saran yang membangun dari berbagai pihak sangat kami harapakan demi perbaikan. Amin Yaa Rabbal ’Alamiin.

Sukabumi, ………….. Mei 2024

Dini Aryani

Penulis

**HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI**

**TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS**

Sebagai sivitas akademik UNIVERSITAS NUSA PUTRA , saya yang bertanda tangan di bawah ini:

Nama : ………………....................................................

NIM : ………………………….....................................

Program Studi : .............................................................................

Jenis karya: Tugas Akhir

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Nusa Putra **Hak Bebas Royalti Noneksklusif *(Non-exclusive Royalty- Free Right*)** atas karya ilmiah saya yang berjudul :

………………………................................................................................................

....................................................................................................................................

....................................................................................................................................

..................................................................................................................................

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Nusa Putra berhak menyimpan, mengalihmedia/format- kan, mengelola dalam bentuk pangkalan data (database), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : …………………….

Pada tanggal : …………………….

Yang menyatakan

………………………………….

# DAFTAR ISI

# DAFTAR TABEL

# DAFTAR GAMBAR

# DAFTAR LAMPIRAN

# BAB I PENDAHULUAN

## 1.1 Latar Belakang

Pada tanggal 14 Februari 2024, Indonesia menyelenggarakan pemilihan umum serentak untuk menentukan pemimpin serta anggota legislatif di tingkat kabupaten/kota, provinsi, nasional, dan juga wakil-wakil di DPD-RI. Dari sekitar 204,8 juta pemilih tetap yang terdaftar, lebih dari 116,5 juta atau sekitar 56 persen adalah generasi Z dan milenial. Generasi Z, kelahiran tahun 1995 hingga 2000-an, memiliki hak pilih sekitar 22,8 persen. Sedangkan milenial, yang lahir antara tahun 1980 hingga 1994, memiliki suara sekitar 33,6 persen dari total [1]. Hal tersebut menunjukkan bahwa pemilu 2024 akan diwarnai oleh partisipasi yang dominan dari kalangan pemilih muda yaitu generasi Z dan Milenial.

Kedua generasi ini telah tumbuh dalam era digital, di mana akses ke *internet* dan media sosial menjadi bagian penting dari kehidupan sehari-hari generasi tersebut. Dukungan dan partisipasi politik generasi milenial ini akan terbentuk jika terus menerus terpapar media digital yang berisikan sosialisasi atau informasi/berita tentang kampanye mendukung kebijakan-kebijakan pemerintah yang positif. Mengingat generasi ini adalah mayoritas pengguna media digital (*internet*, media sosial dan lain sebagainya) [2].

Generasi Z dan Milenial aktif dalam berdiskusi, berbagi artikel berita, dan mencari opini dari berbagai sumber yang tersedia secara daring. Dengan demikian, *internet* menjadi sumber utama bagi generasi milenial dan Z untuk membentuk pandangan politik dan membuat keputusan dalam pemilihan umum. Interaksi ini kemudian dipermudah dengan adanya media massa *online* yang menyajikan berita-berita aktual yang bisa diakses semua kalangan. Ketersediaan dan kemudahan dalam mengakses media massa secara *online* membuat media ini menjadi pilihan yang sangat diminati masyarakat [3].

Detik.com adalah situs berita daring di Indonesia yang didirikan pada 9 Juli 1998 oleh Budiono Darsono, Yayan Sopyan, Abdul Rahman, dan Didi Nugrahadi. Sejak tanggal 3 Agustus 2011, detik.com menjadi bagian dari PT Trans Corporation, anak perusahaan CT Corp. Nama "detik.com" diambil dari tabloid DeTik yang didirikan pada tahun 1977. Awalnya fokus pada politik, ekonomi, dan teknologi informasi, tetapi kemudian juga menyajikan berita hiburan dan olahraga. Detik.com dikenal dengan pembaruan berita secara cepat dan merupakan situs informasi digital paling populer di Indonesia. Menurut data dari *SimilarWeb* pada Januari 2024, Detik.com menduduki peringkat ke 3 teratas sebagai penerbit berita dan media terkemuka.

Detik.com secara aktif memberitakan mengenai pemilu 2024, terbukti dengan adanya sub kanal khusus yang memberitakan pemilu. Hal ini menyebabkan melimpahnya informasi pada media tersebut. Tidak seperti sub kanal pemilu lainnya (seperti sub kanal pemilu Kompas dan CNN), pada sub kanal pemilu Detik belum ada pengkategorian berita berdasarkan topik tertentu. Hal ini menyulitkan pembaca untuk mencari informasi yang lebih mendalam pada sub kanal tersebut. Namun, melacak dan menganalisis topik-topik yang muncul dalam cakupan berita menjadi tantangan, terutama mengingat volume besar artikel berita yang dipublikasikan setiap hari. Oleh karena itu, diperlukan pendekatan yang efisien dan efektif untuk memodelkan topik-topik terkait Pemilu 2024 dari artikel berita detik.

*Topic* *Modelling* adalah salah satu teknik di dalam *Natural* *Language* *Processing* (NLP) untuk menganalisis teks [4]. Algoritma *topic* *modelling* digunakan untuk mengidentifikasi pola tersembunyi dalam rangkaian kata yang terdistribusi di dalam sebuah kumpulan dokumen. Hasil dari teknik ini adalah sekumpulan topik yang terdiri dari beberapa kelompok kata yang muncul bersamaan dalam dokumen berdasarkan pola tertentu. Penelitian ini akan menggunakan Meta Deskripsi sebagai pengukurannya. Meta Deskripsi adalah keterangan yang digunakan untuk menjelaskan apa maksud dari konten yang dibuat [5].

Pada penelitian sebelumnya yang membandingkan beberapa metode *Natural* *Language* *Processing*, *Bidirectional* *Encoder* *Representations* *Transformers* (BERT) dianggap lebih baik dibandingkan model lainnya, termasuk LDA [6]. BERT adalah sad model representasi bahasa baru yang dirancang untuk dapat melakukan pelatihan representasi dua arah dari teks yang tidak berlabel [7]. Kemudian, salah satu yang pustaka yang memanfaatkan BERT sebagai model untuk pemodelan topik adalah BERTopic [8]. BERTopic merupakan model transformer BERT yang dikhususkan untuk pemodelan topik dengan menggunakan pendekatan c-TF-IDF. Sehingga pemodelan topik dapat diinpretasikan dengan mudah tanpa mengurangi kata – kata kunci dalam deskripsi kata yang dilakukan topik modeling [9]

Untuk itu, penulis merasa tertarik untuk melakukan penelitian dengan judul Pemodelan Topik Pemilu 2024 Menggunakan Metode BERTopic pada Artikel Berita Detik.com

## 1.2 Rumusan Masalah

Rumusan masalah yang sesuai untuk penelitian ini dapat dirumuskan sebagai berikut:

1. Bagaimana menerapkan BERTopic untuk pemodelan topik pada meta deskripsi artikel berita terkait Pemilu 2024?
2. Apa saja topik-topik utama yang mendominasi cakupan berita Pemilu 2024 pada portal berita Detik.com?

## 1.3 Batasan Masalah

Dengan merujuk pada konteks masalah yang telah diuraikan sebelumnya, penulis memberikan batasan penulisan untuk memastikan bahwa pembahasan dan isi tulisan tetap sesuai dengan judul penelitian. Batasan-batasan yang diberlakukan melibatkan hal-hal berikut:

1. Data yang dikumpulkan adalah data dalam Bahasa Indonesia.
2. Data yang digunakan merupakan data meta deskripsi artikel berita pada sub kanal pemilu dari tanggal 1 September 2023 hingga 14 Februari 2024 pada portal berita Detik.com.
3. Menggunakan bahasa pemrograman Python.
4. Penggunaan pustaka BERTopic untuk pemodelan topik

## 1.4 Tujuan

Tujuan yang hendak dicapai dalam penelitian ini adalah:

1. Menerapkan metode BERTopic untuk memodelkan topik pada meta deskripsi artikel berita Pemilu 2024.
2. Mengidentifikasi dan menganalisis topik-topik utama yang mendominasi cakupan berita Pemilu 2024 di portal berita Detik.com.
3. Memberikan informasi mengenai tren topik pemilu 2024 pada artikel berita di portal berita Detik.com.
4. Mengelompokkan berita Pemilu 2024 dengan topik yang serupa untuk mempermudah pemahaman.

## 1.5 Manfaat Penelitian

Manfaat dari penelitian ini di antaranya sebagai berikut:

1. Manfaat Bagi Penulis
2. Memberikan pengetahuan lebih kepada penulis dalam menganalisis *topic* *modelling* terhadap meta deskripsi berita pemilu 2024 menggunakan BERTopic.
3. Sebagai salah satu syarat dalam penyelesaian gelar strata (S1) program studi Teknik Informatika Universitas Nusa Putra.
4. Manfaat Bagi Universitas
5. Menjadi sebuah tolak ukur bagi Universitas dalam menentukan keberhasilan dan kemampuan penulis dalam mengimplementasikan ilmu yang sudah didapatkan selama menempuh pendidikan perkuliahan di universitas.
6. Sebagai referensi untuk penelitian selanjutnya.
7. Manfaat Bagi Masyarakat
8. Informasi mengenai hasil *topic* *modelling* terhadap topik pemilu 2024.
9. Dapat dijadikan studi literatur untuk penelitian selanjutnya.

## 1.6 Sistematika Penulisan

Sistematika penulisan dalam tugas akhir ini terdiri dari beberapa bab yang dirancang secara terstruktur untuk menyajikan isi penelitian dengan jelas. Berikut adalah gambaran garis besar mengenai konten masing-masing bab:

### **BAB I: PENDAHULUAN**

Merupakan bab awal yang memberikan latar belakang, permasalahan, tujuan, manfaat, dan batasan penelitian.

### **BAB II: TINJUAN PUSTAKA**

Berisi *review* literatur dan penelitian terkait untuk membangun landasan teoritis dan kerangka konseptual.

*Top of Form*

### **BAB III: METODE PENELITIAN**

Pada bab 3, yaitu metodologi penelitian ini membahas mengenai metode serta langkah-langkah yang digunakan penulis dalam melakukan penelitian yang dilakukan. Bab ini mencakup objek penelitian, analisis, tahap penelitian, alat dan bahan serta perancangan sistem.

### **BAB IV: HASIL**

Bab ini mencakup penjelasan tentang penerapan hasil dan evaluasi dari uji coba penelitian mulai dari tahap analisis, perancangan, implementasi perancangan, hasil pengujian, dan penerapannya. Penjelasan ini mencakup aspek teoritis baik secara kualitatif, kuantitatif, maupun statistik.

### **BAB V: PENUTUP**

Bab ini merupakan penutup dari laporan atau tesis yang menyajikan rangkuman dari keseluruhan penelitian serta memberikan kesimpulan, implikasi praktis, saran untuk penelitian masa depan, dan pesan terakhir dari penulis.

# BAB II TINJAUAN PUSTAKA

## 2.1 Penelitian Terkait

Dalam penyusunan skripsi ini, penulis sedikit banyak terinspirasi dan merujuk pada penelitian sebelumnya yang berkaitan dengan latar belakang masalah pada skripsi ini. Berikut ini penelitian terdahulu yang berhubungan dengan skripsi ini antara lain :

Tabel 2.1 Penelitian terkait

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **No** | **Judul** | **Penulis dan**  **Tahun** | **Ringkasan** | | | **Referensi** |
| 1 | Komparasi K- Means *Clustering* dan Latent Dirichlet Allocation Pada Kasus Short Text *Topic* *Modelling*  (Studi Kasus : Data Judul Berita Pemilu 2024) | Enny Fitriani Nasution (2023) | Penelitian membandingkan model *K-means* *Clustering* dan *Latent Dirichlet Allocation* (LDA) menggunakan judul berita terkait Pemilu 2024 dari AntaraNews, Kompas, Detik, Viva, dan IDNTimes. Data diperoleh melalui teknik *scraping* dan dianalisis menggunakan *confusion matrix* dan Davies-Bouldin Index. | | | [10] |
|  | Persamaan | Melakukan pemodelan topik berita pemilu 2024 | | | | |
|  | Perbedaan | Model yang digunakan, Portal berita, rentang  waktu pengambilan data, dan implementasi | | | | |
| 2 | *Topic Modelling and Opinion Analysis On Climate Change Twitter*  *Data Using LDA And BERT Model* | Samson Ebenezar Uthirapathy, Domnic Sandanam (2023) | Studi ini bertujuan untuk membandingkan performa model BERT (*Bidirectional* *Encoder* *Representations* *from Transformers*) dan LDA  *(Latent Dirichlet Allocation)* dalam melakukan pemodelan topik dan analisis opini. Hasilnya model BERT *uncased* menunjukkan hasil terbaik seperti *precision* sebesar 91,35%, *recall* sebesar 89,65%, dan *accuracy*  sebesar 93,50% dibandingkan dengan metode LDA. | | | [11] |
|  | Persamaan | Dalam penelitian ini memiliki persamaan yaitu  menggunakan model BERT untuk pemodelan topik. | | | | |
|  | Perbedaan | Topik penelitian | | | | |
| 3 | *BERT for Arabic Topic Modeling: An Experimental Study on*  *BERTopic Technique* | Abeer Abuzayed, Hend Al-Khalifa (2021) | | Penelitian tersebut melakukan perbandingan performa model BERTopic dengan model lainnya yaitu, LDA  *(Latent Dirichlet Allocation)* dan NMF (*Non-Negative Matrix Factorization*) dalam melakukan pemodelan topik untuk bahasa arab. Hasilnya model BERTopic secara keseluruhan memiliki performa yang lebih baik dibandingkan kedua model lainnya. | [12] | |
|  | Persamaan | Menggunakan model BERTopic | | | | |
|  | Perbedaan | Topik penelitian | | | | |
| 4 | Pemodelan Topik Pada Ulasan Hotel Menggunakan Metode BERTopic Dengan Prosedur c-TF-IDF | Mertayasaa I, Darmawana I  (2022) | | Penelitian tersebut melakukan pemodelan topik menggunakan data *review* pengguna terhadap layanan hotel. Hasil pemodelan topik menggunakan metode BERTopic menghasilkan 78 topik dengan nilai *coherence* topik sebesar 0,07287 dan keberagaman topik sebesar 0,496154. Semakin sedikit jumlah topik yang dihasilkan, maka nilai *coherence* dan keberagaman topik akan menurun. | [13] | |
|  | Persamaan | Menggunakan pustaka BERTopic untuk melakukan pemodelan topik | | | | |
|  | Perbedaan | Topik penelitian | | | | |
| 5 | Pemodelan Topik pada Media Berita Online Menggunakan Latent Dirichlet Allocation (Studi Kasus Merek Somethinc) | Puspita E, Shiddieq D, Roji F (2024) | | Penelitian ini mengkaji topik-topik yang sering dibahas dalam berita *online* tentang merek kosmetik lokal Somethinc di Indonesia. Menggunakan metode *topic modeling*, khususnya *Latent Dirichlet Allocation* (LDA), penelitian ini menemukan bahwa topik terbaik terdiri dari 6 topik dengan nilai *coherency* sebesar 0.404. Kesimpulannya, jumlah topik 6 menghasilkan topik-topik yang relevan dan representatif terkait merek *Somethinc*. | [14] | |
|  | Persamaan | Melakukan pemodelan topik media berita *online* | | | | |
|  | Perbedaan | Studi kasus dan model yang digunakan | | | | |

Berdasarkan hasil penelitian terdahulu yang telah disampaikan, dapat diambil beberapa kesimpulan yang relevan untuk penelitian yang akan dilakukan terkait pemodelan topik pemilu 2024 menggunakan metode BERTopic pada artikel berita Detik.com:

1. Pilihan Model: Pemilihan model BERTopic untuk pemodelan topik terkait Pemilu 2024 pada artikel berita Detik.com sangat relevan berdasarkan penemuan terdahulu. BERTopic telah terbukti lebih efisien dalam mengolah data besar seperti judul berita terkait pemilihan dan mampu memberikan hasil yang memadai dalam analisis topik.
2. Performa: Studi sebelumnya menunjukkan bahwa penggunaan model BERT terutama BERTopic untuk pemodelan topik terbukti menghasilkan performa yang lebih baik daripada model lainnya seperti LDA  *(Latent Dirichlet Allocation)* dan NMF (*Non-Negative Matrix Factorization*).
3. Relevansi Temuan: Temuan dari penelitian sebelumnya juga dapat memberikan panduan dalam memperkirakan sumber daya komputasi atau biaya yang dibutuhkan dalam implementasi BERTopic pada skenario analisis topik berita Pemilu 2024 di *platform* Detik.com.

Dengan demikian, penggunaan BERTopic yang dibarengi dengan model BERT dalam penelitian ini diharapkan dapat memberikan hasil yang efisien dan informatif dalam pemodelan topik terkait Pemilu 2024 pada artikel berita Detik.com.

## 2.2 Landasan Teori

### **2.2.1 *Natural* *Language* *Processing***

Pengolahan Bahasa Alami (*Natural* *Language* *Processing*/NLP) merupakan cabang dari ilmu kecerdasan buatan (*Artificial Intelligence*/AI) yang memungkinkan komputer untuk mengerti dan memproses teks manusia dari beragam bahasa [15]. Naufal menjelaskan bahwa *Natural Language Processing* (NLP) adalah suatu sistem yang mempermudah interaksi manusia dengan bahasa mesin, di mana bahasa pemrograman tersebut diolah sehingga dapat dimengerti oleh manusia secara lebih mudah [16]. Ini melibatkan pengembangan teknologi dan algoritma yang memungkinkan komputer untuk memproses dan memahami teks dan ucapan manusia dalam berbagai bahasa, serta menghasilkan *output* yang bermanfaat berdasarkan pemahaman tersebut. Salah satu implementasi NLP yaitu pemodelan topik atau *topic* *modelling*.

### **2.2.2 *Topic* *Modelling***

*Topic* *modelling* adalah suatu pendekatan dalam analisis teks yang bertujuan untuk mengidentifikasi dan mengekstraksi topik atau tema yang tersembunyi dalam kumpulan dokumen. Tujuan utama dari *topic* *modelling* adalah untuk mengelompokkan dokumen ke dalam kategori atau topik tertentu berdasarkan kontennya [17]. *Topic* *modelling* adalah sebuah metode analisis data yang digunakan untuk mengidentifikasi topik-topik yang muncul secara alami dalam suatu koleksi dokumen atau teks [18]. Metode ini sering digunakan untuk mengatasi volume besar dokumen teks dan membantu pemahaman terhadap isu-isu atau topik utama yang dibahas dalam koleksi dokumen tersebut. Pemodelan topik masuk ke dalam kategori *unsupervised* *learning* pada pembelajaran mesin di mana tidak ada label atau definisi teks yang dijadikan acuan, dan jumlah topik yang akan dipelajari ditentukan sebelum proses pelatihan dimulai [19].

### **2.2.3 *Clustering***

*Clustering* mengacu pada salah satu metode pembelajaran mesin yaitu *unsupervised* *learning*. Dalam implementasi metode *unsupervised* *learning* pelatihan model dilakukan tanpa memberikan pengetahuan mengenai data yang digunakan (label) pada *dataset* yang digunakan. Model *unsupervised* *learning* hanya menarik kesimpulan dari *dataset* berdasarkan kedekatan data atau *clustering* [20].

*Clustering* merupakan metode untuk mengelompokkan data ke dalam kategori-kategori yang spesifik [21]. Klastering (*clustering*) adalah suatu teknik dalam analisis data yang bertujuan untuk mengelompokkan objek atau data ke dalam kelompok atau klaster berdasarkan kesamaan karakteristik tertentu. Dalam konteks ini, objek atau data yang serupa akan dikelompokkan bersama, sementara objek atau data yang berbeda akan berada di kelompok yang berbeda [22]. Menurut M. S. Yang dan K. P. Sinaga, Tujuan utama dari klastering adalah mencari struktur tersembunyi dalam data dan membentuk kelompok yang bermakna tanpa adanya label kelas sebelumnya [23]. Klastering digunakan dalam berbagai bidang seperti analisis data, pengelompokan konsumen, pengelompokan dokumen teks, pengelompokan citra, dan banyak lagi.

### **2.2.4 Artikel Berita**

Berdasarkan definisi dari Kamus Besar Bahasa Indonesia (KBBI), berita merupakan narasi atau informasi tentang suatu peristiwa atau kejadian yang menarik perhatian. Dalam sebuah laporan berita, terdapat unsur fakta yang menggambarkan kejadian secara obyektif, serta opini yang ditampilkan dengan cara menarik bagi masyarakat. Namun, tidak semua fakta dan pendapat dapat dipilih oleh media untuk dijadikan sebagai berita [24]. Di era modern ini, metode penyampaian berita yang sebelumnya dikenal melalui surat kabar, majalah, radio, dan televisi, telah mengalami perubahan signifikan dengan beralih ke *platform* berbasis *internet* atau diubah menjadi artikel *online* [25].

Artikel berita adalah suatu bentuk tulisan jurnalistik yang memberikan informasi tentang peristiwa atau kejadian terbaru. Artikel berita bertujuan menyampaikan fakta dan informasi yang relevan kepada pembaca agar mereka dapat memahami apa yang terjadi di sekitar mereka. Artikel berita dapat mencakup berbagai topik seperti politik, ekonomi, olahraga, hiburan, dan sebagainya.

1. Judul

Judul adalah frasa atau kalimat pendek yang digunakan untuk merangkum atau menggambarkan isi dari suatu karya, seperti artikel, buku, atau *posting* *online*. Judul bertujuan untuk menarik perhatian pembaca dan memberikan gambaran singkat tentang topik atau konten yang akan dibahas.

1. Meta Deskripsi

Meta deskripsi, atau *meta description*, adalah deskripsi singkat yang ditempatkan di *tag* HTML pada halaman *web.* Deskripsi ini bertujuan memberikan gambaran tentang isi halaman *web* dan muncul sebagai cuplikan teks dalam hasil pencarian mesin pencari. Meta deskripsi membantu pembaca dan mesin pencari memahami konten halaman tanpa harus membukanya.

### **2.2.5 *Bidirectional* *Encoder* *Representations* *Transformers* (BERT)**

BERT (*Bidirectional* *Encoder* *Representations* *from Transformers*) adalah model bahasa yang dikembangkan oleh Google yang menggunakan arsitektur Transformer untuk memahami konteks kata - kata dalam suatu teks. BERT dirancang untuk menghasilkan representasi kata yang sangat kontekstual dan dapat digunakan dalam berbagai tugas pemrosesan bahasa alami (NLP) [26].

Berikut beberapa kelebihan model BERT jika dibandingkan dengan model lainnya [27]:

1. *Pre-trained model*:

BERT memiliki banyak variasi *pre-trained model* dari berbagai jenis bahasa. Sehingga pengguna tidak perlu melakukan pelatihan model dari awal.

1. Representasi berbasis *fine-tuning*:

BERT merupakan model pertama yang menggunakan *fine-tuning* untuk mencapai performa yang sangat baik dalam pemahaman kalimat dan token dalam skala yang besar. BERT dapat digunakan dalam berbagai tugas dengan arsitektur khusus yang lebih spesifik.

1. *Encoder* sebagai *input*:

BERT menggunakan *encoder* sebagai *input*, tidak secara sekuensial yang memungkinkannya untuk memahami hubungan kontekstual dari setiap token dengan baik. Hal tersebut dapat tercapai dikarenakan posisi *encoding* BERT memungkinkan untuk memproses *input* yang panjang.

1. Akses pada setiap token:

BERT memiliki akses dengan setiap token, yang memungkinkan BERT dapat mengerti makna dari setiap token secara bersamaan dari dua arah sekaligus.

### **2.2.6 *Scraping***

*Web Scraping* merupakan metode untuk mengambil informasi dari halaman *web.* Proses *web scraping* erat kaitannya dengan *indexing* *web.* Ada beberapa langkah dalam mengembangkan teknik *web scraping*. Pertama, pengembang program mempelajari struktur dokumen HTML dari situs *web* yang akan diambil informasinya. Tujuannya adalah untuk mengidentifikasi dan menangkap informasi yang diinginkan. Selanjutnya, pengembang program mempelajari teknik navigasi pada situs *web* tersebut untuk dapat mengikuti pola yang sama pada aplikasi *web scraping* yang sedang dikembangkan [28].

Proses *web scraping* pengambilan data dari *Internet* dapat dibagi menjadi dua tahap yang berurutan, yakni mengambil sumber daya *web* terlebih dahulu, kemudian mengekstraksi informasi yang diinginkan dari data yang telah didapat [29]. Dalam konteks penelitian ini *web scraping* dilakukan untuk mendapatkan artikel berita pemilu dari portal berita Detik.com menggunakan rentang waktu tertentu.

### **2.2.7 *Preprocessing***

*Preprocessing* dilakukan untuk menghapus gangguan suara dan mengubah bentuk kata agar kosakata menjadi lebih seragam, sehingga volume kosakata dapat dikurangi [30]. Menurut Tu *preprocessing* teks bertujuan untuk mengubah data yang tak terstruktur menjadi lebih terstruktur [31].

Dapat disimpulkan jika *preprocessing* merujuk pada serangkaian langkah atau proses yang dilakukan pada data mentah sebelum data tersebut dapat digunakan untuk analisis atau pemrosesan lebih lanjut. Tujuan dari *preprocessing* adalah untuk membersihkan, mengorganisir, dan mempersiapkan data agar menjadi lebih sesuai untuk analisis atau penggunaan tertentu. Pada penelitian ini *preprocessing* dilakukan pada data teks (text *preprocessing*). Jenis text *preprocessing* yang dapat diimplementasikan yaitu [32][33]:

1. *Case folding*

*Case folding* mengacu pada proses mengubah semua huruf dalam teks menjadi huruf kecil (*lowercase*) atau huruf besar (*uppercase*), tergantung pada kebutuhan analisis. Hal ini dilakukan untuk mengurangi variasi yang mungkin timbul akibat perbedaan kasus dalam kata-kata yang sama. Misalnya, "Data", "data", dan "DATA" akan diubah menjadi "data".

1. *Cleaning*

Tahap pembersihan (*cleaning*) mencakup serangkaian langkah untuk menghilangkan karakter atau elemen yang tidak relevan atau mengganggu dalam teks. Ini bisa meliputi menghapus tanda baca, karakter spesial, angka, atau tautan *web* yang tidak diinginkan.

1. *Tokenizing*

*Tokenizing* atau tokenisasi adalah proses membagi teks menjadi unit-unit yang lebih kecil yang disebut token. Token bisa berupa kata-kata, frasa, atau simbol tergantung pada aturan yang digunakan. Biasanya, tokenisasi dilakukan dengan memisahkan teks berdasarkan spasi antar-kata, namun bisa juga melibatkan aturan tambahan seperti pemisahan berdasarkan tanda baca atau karakter tertentu. Hasil tokenisasi akan menghasilkan kumpulan token yang dapat dijadikan dasar untuk analisis teks lebih lanjut, seperti klasifikasi, analisis sentimen, atau pemodelan bahasa.

1. *Stopword removal*

*Stopword removal* merupakan proses penghapusan kata – kata umum seperti kata hubung dan kata bantu yang dianggap tidak penting dalam pemrosesan data teks. Penghapusan tersebut dapat mereduksi dimensi dari data yang diproses, sehingga dapat mengurangi waktu pemrosesan data oleh modeling.

### **BERTopic**

BERTopic merupakan pustaka yang menerapkan model BERT untuk keperluan spesifik yaitu pemodelan topik [8]. BERTopic adalah teknik pemodelan topik yang menggunakan *embedding* BERT dan menunjukkan bahwa tugas tersebut dapat ditangani secara efektif melalui pendekatan berbasis *clustering* [34]. BERTopic adalah sebuah model topik yang memperluas prosesnya dengan mengekstraksi representasi topik yang koheren melalui pengembangan variasi berbasis kelas dari TF-IDF. Lebih spesifik lagi, BERTopic menghasilkan *embedding* dokumen dengan model bahasa berbasis *transformer* yang telah dilatih sebelumnya, mengelompokkan *embedding* ini, dan akhirnya menghasilkan representasi topik dengan prosedur TF-IDF berbasis kelas (c-TF-IDF) [35].

Sebelum dilakukan implementasi, dataset yang digunakan untuk BERTopic memerlukan *preprocessing* manual, seperti *tokenizing*, penghapusan *stopword*, *case folding*, *stemming* dan *lemmatization*, untuk menghilangkan informasi tambahan dan mengurangi kompleksitas teks [36]. Dalam pemodelan topik, BERTopic melewati 5 proses utama ditambah 1 proses opsional yang ditunjukkan pada Gambar 2.1 sebagai berikut [9]:



Gambar 2.1 Proses BERTopic

1. *Embeddings*

BERTopic dimulai dengan mengubah dokumen *input* kita menjadi representasi numerik. Meskipun ada banyak cara untuk mencapainya, BERTopic menggunakan *sentence-transformers* "all-MiniLM-L6-v2" untuk bahas inggris dan “paraphrase-multilingual-MiniLM-L12-v2” untuk multibahasa karena mampu menangkap kesamaan semantik antara dokumen.

1. *Dimensionality Reduction*

Salah satu proses penting dari BERTopic adalah pengurangan dimensi dari *embedding input*. Karena *embedding* seringkali memiliki dimensi yang tinggi yang dapat mempersulit *clustering* atau pengelompokan topik. UMAP (*Uniform Manifold Approximation and Projection*) digunakan sebagai standar dalam BERTopic karena dapat menangkap baik ruang dimensi tinggi lokal maupun *global* ke dalam dimensi yang lebih rendah.

1. *Clustering*

Setelah mengurangi dimensi dari *embedding* *input*, dilakukan proses *clustering* atau mengelompokkan *embedding* *input* ke dalam kelompok-kelompok *embedding* yang mirip untuk mengekstrak topik-topik pada *dataset*. Dalam BERTopic, standar yang digunakan yaitu menggunakan HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) karena mampu menangkap struktur dengan kepadatan yang berbeda. Namun pada pustaka BERTopic dapat dilakukan penyesuaian model clustering lainnya sesuai kebutuhan, seperti menggunakan *K-means*, Agglomerative Clustering, cuML HDBSCAN dan lain sebagainya.

1. *Tokenizer*

Dalam pemodelan topik, kualitas representasi topik sangat penting untuk menginterpretasikan topik, menyampaikan hasil, dan memahami pola-pola. Dalam BERTopic, standar yang digunakan yaitu menggunakan metode CountVectorizer. Metode tersebut digunakan untuk mengonversi koleksi teks menjadi representasi numerik berdasarkan hitungan frekuensi kemunculan setiap kata dalam teks.

1. *Weighting Scheme*

Dalam BERTopic, untuk mendapatkan representasi topik yang akurat dari *matriks bag-of-words* dilakukan pembobotan menggunakan metode TF-IDF. TF-IDF yang digunakan disesuaikan untuk bekerja pada tingkat klaster/kategori/topik daripada tingkat dokumen. Representasi TF-IDF yang disesuaikan ini disebut c-TF-IDF dengan memperhitungkan apa yang membuat dokumen dalam satu klaster berbeda dari dokumen dalam klaster lainnya. Persamaan c-TF-IDF ditunjukkan sebagai berikut.

(2.1)

Keterangan :

: Frekuensi kata X di kelas X

: Frekuensi kata X di semua kelas

: Jumlah rata-rata kata per kelas

1. *Fine-tune Representations (Optional)*

Setelah menghasilkan representasi c-TF-IDF, kita memiliki sekelompok kata yang menggambarkan sebuah koleksi dokumen. C-TF-IDF adalah metode yang dapat dengan cepat menghasilkan representasi topik yang akurat. Menggunakan data tersebut dapat dilakukan penyesuaian topik c-TF-IDF ini menggunakan GPT, T5, KeyBERT, Spacy, dan teknik lainnya.

### ***K-means***

*K-means* merupakan merupakan salah satu algoritma clustering yang dapat melakukan pengelompokan data yang cukup besar secara cepat dan efisien [37]. *K-means* melakukan clustering dengan cara meminimalisir jumlah kuadrat jarak dari setiap titik data dan mendekat ke titik pusat *cluster* data tersebut. Ilustrasi dari klasterisasi menggunakan algoritma *K-means* ditunjukkan pada Gambar 2.2 sebagai berikut [38].

A diagram of a graph

Description automatically generated with medium confidence

Gambar 2.2 Ilustrasi klasterisasi algoritma K-means

Berdasarkan ilustrasi klasterisasi algoritma *K-means* seperti yang ditunjukkan pada Gambar 2.2, dapat diketahui jika algoritma tersebut melakukan pengelompokan berdasarkan kemiripan dari data. Tahapan algoritma *K-means* dalam melakukan clustering dijelaskan sebagai berikut [39].

1. Menentukan jumlah *cluster* atau k;
2. Menentukan pusat *cluster* (centroid) sama dengan jumlah jumlah cluster;
3. Menggunakan persamaan *euclidean distance* atau persamaan lainnya untuk menghitung jarak data ke pusat cluster. Persamaan *euclidean distance* ditunjukkan sebagai berikut.

(2.1)

Keterangan :

: Jarak data ke-x dengan pusat *cluster* ke-y

: Jumlah data

: Index atribut, (i = 1, 2, 3, 4, ........, n)

: Atribut data ke-i

: Atribut pusat *cluster* ke-i

1. Tahap selanjutnya adalah memperbaharui titik pusat *cluster* dan mengulangi tahap perhitungan jarak hingga nilai dari titik pusat tidak berpindah. Persamaan untuk memperbaharui titik pusat *cluster* ditunjukkan sebagai berikut.

(2.1)

Keterangan :

: Titik pusat cluster

: Jumlah data

: Index atribut, (i = 1, 2, 3, 4, ........, n)

: data ke-i

Penentuan jumlah *cluster* atau k pada penelitian ini dilakukan menggunakan metode *Elbow* untuk menjadi jumlah *cluster* optimal dari data yang digunakan [40].

### **Evaluasi**

#### **2.2.9.1 Topic Coherence**

Dalam *topic* *modelling* evaluasi dilakukan proses evaluasi dilakukan dengan mengukur kecocokan makna antara berbagai aspek kata dalam suatu topik tertentu, yang sering disebut sebagai koherensi topik atau *topic* *coherence* [41]. Evaluasi tersebut menjadi penting dalam konteks pemodelan topik untuk mengukur seberapa baik kata-kata yang dipilih mencerminkan topik yang sedang dibahas dan seberapa konsisten hubungan antara kata-kata tersebut dalam konteks topik tersebut.

*Topic coherence* adalah suatu proses yang menerima topik dan korpus referensi sebagai *input*-nya dan kemudian menghasilkan nilai koherensi dari topik tersebut sebagai *output*-nya. Tahapan-tahapan dari proses ini meliputi segmentasi, perhitungan probabilitas, pengukuran konfirmasi, dan agregasi [42].

1. Segmentasi berfungsi untuk membuat pasangan kata yang digunakan untuk menilai koherensi topik. Rumus untuk melakukan segmentasi sebagai berikut.

(2.2)

1. Perhitungan probabilitas berperan dalam menghitung probabilitas dari setiap pasangan kata. Dalam penelitian ini, digunakan metode 𝑃𝑏𝑑 yang menghitung 𝑃(𝑊’) sebagai jumlah dokumen yang mengandung kata 𝑊’ dibagi dengan jumlah total dokumen, serta 𝑃(𝑊’𝑊\*) sebagai jumlah dokumen yang mengandung kata 𝑊’ dan 𝑊\* dibagi dengan jumlah total dokumen.
2. Pengukuran konfirmasi berperan dalam menghitung nilai konfirmasi dengan memanfaatkan 𝑃(𝑊’) dan 𝑃(𝑊’𝑊\*). Terdapat beberapa rumus yang digunakan untuk menghitung nilai konfirmasi, dan penelitian ini menggunakan rumus tertentu untuk hal tersebut.

(2.3)

1. Agregasi merupakan nilai rata-rata dari seluruh nilai yang diperoleh dari setiap segmentasi.

Keterangan :

: Himpunan pasangan kata

𝑊 : Himpunan kata

𝑊’ : Kata 1

𝑊\* : Kata 2

: Pengukuran konfirmasi dari (𝑊’𝑊\*)

#### **2.2.9.2 *Silhouette Coefficient***

*Silhouette Coefficient* merupakan salah satu metode evaluasi yang populer pada clustering atau pengelompokan. Nilai *Silhouette Coefficient* dapat menggambarkan kualitas dari *cluster* yang dihasilkan [37]. Nilai tersebut digunakan untuk menghitung jarak data pada *cluster* ataupun tetangga *cluster* [43]. Menggunakan hasil tersebut dapat diketahui adapkah model yang akan diterapkan sesuai atau tidak. Perhitungan nilai *Silhouette Coefficient* melewati beberapa tahapan sebagai berikut [44].

1. Tahap pertama untuk menghitung nilai *Silhouette Coefficient* adalah menghitung rata – rata data ke-i pada *cluster* yang sama. Dengan asumsi bahwa data berada di *cluster* A, berikut persamaan untuk perhitungan tersebut.

(2.4)

Keterangan :

A : Jumlah data pada *cluster* A

1. Tahap kedua yaitu menghitung nilai minimum jarak rata – rata data ke-i terhadap semua data pada *cluster* yang berbeda. Dengan asumsi bawa *cluster* berbeda tersebut adalah C, berikut persamaan untuk perhitungan tersebut.

(2.5)

Keterangan :

C : Jumlah data pada *cluster* C

1. Tahap ketiga yaitu menentukan nilai jarak minimum atau . Berikut persamaan untuk menghitung nilai jarak minimum.

(2.6)

1. Setelah nilai dan diketahui maka dapat dihitung untuk nilai *Silhouette Coefficient*. Berikut persamaan untuk perhitungan nilai *Silhouette Coefficient*.

(2.7)

Keterangan :

: *Silhouette Coefficient*

Jika *cluster* B menghasilkan nilai jarak terendah, maka dapat dikatakan sebagai tetangga data ke-i dan *cluster* A merupakan *cluster* terbaik dilanjutkan dengan *cluster* B. Nilai atau *Silhouette Coefficient* berada pada rentang -1 sampai dengan 1. Berikut contoh interpretasi hasil perhitungan nilai *Silhouette Coefficient* untuk *cluster* A dan *cluster* B.

: Data ke-i dikelompokkan baik dalam *cluster* A;

: Data ke-i dikelompokkan di antara *cluster* A dan *cluster* B;

: Data ke-i dikelompokan kuran baik (lebih dekat dengan *cluster* B)

Kemudian secara umum interpretasi hasil evaluasi *Silhouette Coefficient* dapat dijelaskan sebagai berikut [45].

Tabel 2.2 Interpretasi nilai Silhouette Coefficient

|  |  |
| --- | --- |
| **Nilai *Silhouette Coefficient*** | **Interpretasi** |
| 0.71 – 1.00 | Struktur kuat |
| 0.51 – 0.70 | Struktur baik |
| 0.26 – 0.50 | Struktur lemah |
| ≤ 0.25 | Tidak terstruktur |

#### **2.2.9.3 *Elbow***

Metode *Elbow* merupakan salah satu metode evaluasi untuk menganalisis performa hasil klasterisasi, biasanya pada klasterisasi non-hierarki (*partitioning*) seperti *K-means* [46]. Metode *Elbow* atau metode siku merupakan metode yang digunakan untuk menentukan jumlah *cluster* optimal dalam clustering [40]. Metode ini dilakukan dengan menghitung nilai *Sum of Square Error* (SSE) pada setiap *cluster*, dengan memperhatikan besar selisih antar nilai *cluster*. Jumlah *cluster* dikatakan baik jika selisih nilai tersebut membentuk sudut sikut [47]. Sehingga penggunaan metode ini dapat memberikan informasi model optimal yang dapat diterapkan untuk *clustering* data yang digunakan.

Persamaan untuk menghitung nilai *Sum of Square Error* (SSE) ditunjukkan pada persamaan berikut [48].

(2.7)

Keterangan :

: Nilai *Sum of Square Error* : Jumlah *cluster*

: Jumlah data

: Jumlah *cluster*  pada *cluster* ke-

#### **2.2.9.4 *Davies Bouldin Index***

*Davies Bouldin Index* (DBI), yang sering disebut sebagai reliabilitas klasifikasi indeks, merupakan sebuah skema evaluasi *cluster* internal yang menilai keberhasilan klasterisasi berdasarkan kualitas dan kompaknya klaster yang dihasilkan [49]. Semakin rendah nilai DBI, semakin baik pemisahan antar-klaster dan semakin kompak klasternya [50]. Langkah – langkah dalam menentukan nilai DBI adalah sebagai berikut [51].

1. Menghitung pusat klaster untuk setiap klaster dan ukuran dispersi klaster;
2. Menghitung jarak antara pusat-pusat klaster;
3. Menggunakan rumus DBI untuk menentukan nilai R antara klaster;
4. Mengulangi proses untuk setiap pasangan klaster;
5. Menghitung DBI keseluruhan dengan rumus sebagai berikut.

(2.8)

Keterangan :

: *Davies Bouldin Index* secarakeseluruhan : Jumlah total *cluster*

: *Davies Bouldin Index* antara *cluster*  dan

#### **2.2.9.5 *Calinski Harabasz Index***

*Calinski Harabasz Index* (CHI) merupakan metrik yang menghitung perbandingan antara dispersi antar kluster dan dispersi di dalam kluster, diukur sebagai jumlah kuadrat jarak, untuk seluruh kluster atau secara sederhana mengukur rasio antara varians dalam klaster dengan varians antar klaster [52]. Nilai indeks yang tinggi menunjukkan bahwa klaster-kelompok data lebih terpisah satu sama lain [53]. Langkah – langkah perhitungan CHI adalah sebagai berikut [54].

1. Melakukan perhitungan dan menggunakan persamaan berikut.

(2.9)

(2.10)

Keterangan :

: Banyak data

: *Cluster* ke-

: Titik pusat *cluster* ke-

: Titik ke- dalam *cluster* ke-

: Jumlah titik pada *cluster* ke-

1. Menghitung nilai *Sum of Square Beetween-cluster* (SSB) menggunakan persamaan berikut.

(2.11)

Keterangan :

: Banyak data

: Jumlah cluster

: Titik pusat *cluster* ke-

: Titik ke-i dalam *cluster* ke-

: Jumlah titik pada *cluster* ke-

1. Menghitung nilai *Calinski Harabasz Index* (CHI) menggunakan persamaan berikut.

(2.12)

Keterangan :

: Jumlah cluster

: *Cluster* ke-

: Titik pusat *cluster* ke-

: Titik ke-i dalam *cluster* ke-

1. Menghitung nilai *Sum of Square Within-cluster (SSW)* menggunakan persamaan berikut.

(2.13)

Keterangan :

: Jumlah *cluster*

: Banyak data

: Nilai *Sum of Square Within-cluster*

: Nilai *Sum of Square Beetween-cluster*

: Nilai *Calinski Harabasz Index*

Semakin kecil nilai SSW menunjukkan kedekatan hubungan dalam *cluster*, semakin besar nilai SSB menunjukkan tingkat dispersi yang tinggi, dan semakin tinggi nilai CHI menunjukkan efek *clustering* yang lebih baik.

#### **2.2.9.6 *Dunn Index***

*Dunn Index* (DI) merupakan metrik yang mengukur rasio antara jarak minimum antara dua klaster dengan jarak maksimum dalam klaster-kelompok data. Nilai indeks yang tinggi menunjukkan bahwa klaster-kelompok data lebih terpisah secara jelas, sehingga hasil tersebut semakin baik[17]. Untuk menghitung nilai DI dapat dilakukan menggunakan rumus berikut [18].

(2.13)

Keterangan :

: *Dunn Index*

: Jarak antara *cluster*  dan

: Jarak dalam *cluster*

### **2.2.10 Streamlit**

Streamlit merupakan *framework* yang ditulis dalam bahasa Python dan termasuk dalam kategori dasbor data. *Framework* ini dirancang untuk mempermudah penyajian data dan hasil dalam bentuk *dashboard* *website.* Kelebihan *framework* ini adalah pengguna tidak perlu melakukan konfigurasi *interface* secara langsung pada HTML, CSS, maupun JavaScript karena konfigurasi sudah dibuat dalam fungsi – fungsi yang lebih sederhana [55]. Penggunaan *framework* ini sangat cocok untuk pengguna yang hanya membutuhkan dokumentasi *project* *data* *science* seperti *topic modeling*, dikarenakan konfigurasi yang mudah dan cepat [56].

## 2.3 Kerangka Pemikiran

A screenshot of a black and white screen

Description automatically generated

Gambar 2.3 Kerangka berpikir

Kerangka berpikir penelitian dengan judul “Pemodelan Topik Pemilu 2024 Menggunakan Metode BERTopic pada Artikel Berita Detik.com” dimulai dengan identifikasi masalah yang meliputi dominasi Generasi Z dan Milenial dalam pesta demokrasi serta pencarian informasi mengenai pemilu 2024 melalui artikel berita di portal Detik.com. Selain itu, ditemukan juga masalah untuk mengetahui kualitas dan akurasi berita pemilu 2024, pemahaman mendalam tentang isu-isu pemilu, dan belum adanya pemodelan topik berita pemilu 2024 menggunakan BERTopic pada portal Detik.com.

Dari hasil identifikasi masalah dapat dibuat hipotesis yaitu melakukan pemodelan topik menggunakan model BERTopic pada artikel berita pemilu 2024 di portal Detik.com. Dengan demikian, penelitian ini bertujuan untuk memberikan pemahaman yang lebih mendalam terhadap topik-topik dominan dalam liputan media massa, terutama pada portal Detik.com, dengan fokus pada isu-isu pemilu yang paling mencolok.

Diharapkan bahwa hasil penelitian ini akan memberikan kontribusi dalam meningkatkan pemahaman masyarakat, khususnya Generasi Z dan Milenial, mengenai isu-isu yang berkaitan dengan pemilu 2024. Dengan memodelkan topik-topik berita menggunakan BERTopic, diharapkan kualitas dan akurasi berita pemilu 2024 pada portal Detik.com dapat ditingkatkan. Selain itu, pemodelan topik ini juga diharapkan dapat menjadi landasan untuk pengembangan metode analisis berita yang lebih canggih dan efisien di masa mendatang, serta dapat menjadi referensi bagi penelitian-penelitian selanjutnya dalam bidang analisis berita dan pemilu menggunakan teknologi NLP (*Natural Language Processing*).

# BAB III METODOLOGI PENELITIAN

Dalam penelitian implementasi metode BERT (*Bidirectional Encoder Representations Transformers*) untuk pemodelan topik pemilu 2024 pada artikel berita Detik.com menggunakan pustaka BERTopic, objek yang diteliti merupakan artikel berita terkait pemilu 2024 pada portal berita. Pengumpulan data ini dilakukan melalui data meta deskripsi artikel berita *online*, dalam kasus penelitian ini yaitu pada portal berita Detik.com. Data yang digunakan merupakan data hasil *scraping* terhadap meta deskripsi artikel berita pada sub kanal pemilu dari tanggal 1 September 2023 hingga 14 Februari 2024. Pemodelan topik pemilu 2024 pada artikel berita Detik.com menggunakan metode BERTopic melewati beberapa tahapan yang ditunjukkan pada Gambar 3.1 sebagai berikut:

A close-up of a black and white label

Description automatically generated

Gambar 3.1 Tahap penelitian

## 3.1 Identifikasi Masalah

Identifikasi masalah dilakukan dengan studi literatur. Setelah dilakukan studi literatur didapatkan permasalahan bahwa pada pemilu sebelumnya yaitu tahun 2019 terdapat kecenderungan salah satu portal berita terhadap pihak politik. Sehingga penting dilakukan penelitian terhadap berita terkait pemilu 2024. Kemudian dilakukan penetapan metode penelitian yaitu metode kualitatif.

Metode pada penelitian ini bersifat kualitatif dan menggunakan teknik *Topic* *Modelling*, menggunakan model BERTopic, untuk menganalisis topik-topik dalam artikel berita *online* terkait Pemilihan Umum (Pemilu) Indonesia 2024. BERTopic, sebagai algoritma *Natural* *Language* *Processing* (NLP), digunakan untuk mengidentifikasi pola tersembunyi dalam teks dari sumber berita terpercaya yaitu Detik.com. Pendekatan ini bertujuan memberikan pemahaman mendalam terhadap topik-topik yang dominan dalam liputan media massa terutama pada portal berita Detik.com, dengan fokus pada isu-isu pemilu yang paling mencolok.

## 3.2 Pengumpulan Data

### **3.2.1 Studi Literatur**

Studi literatur dilakukan dengan mencari informasi sumber referensi melalui internet seperti buku, jurnal, maupun artikel lainnya terkait *topic modelling*, *clustering*, BERT dan topik lainnya yang berkaitan dengan penelitian yang dilakukan. Rujukan kepada karya-karya sebelumnya dalam bentuk studi literatur digunakan sebagai panduan atau dasar untuk menyusun penelitian yang direncanakan.

### **3.2.2 *Scraping***

Pengumpulan data meta berita dilakukan dengan metode *scraping* menggunakan *web scraper* yang dibuat menggunakan pustaka BeautifulSoup. Dengan menggunakan *web scraper* tersebut dapat dilakukan pengambilan data meta deskripsi artikel berita pada sub kanal pemilu di Detik.com. Dalam melakukan proses *scraping* penulis membaginya menjadi 4 tahapan sebelum menghasilkan data *preprocessing* secara utuh yaitu *scraping link index*, *scraping* tahap pertama, *scraping* tahap kedua dan penggabungan data *scraping*. Berikut penjelasan masing – masing tahapan yang dilakukan.

1. *Scraping link index*

*Scraping link index* merupakan tahapan untuk mendapatkan *link* atau tautan setiap artikel berita pemilu 2024 di sub kanal pemilu di Detik.com untuk setiap tanggal mulai dari tanggal dari tanggal 1 September 2023 hingga 14 Februari 2024. Tautan tersebut kemudian digunakan untuk mengakses meta deskripsi setiap artikel yang ada.

1. *Scraping* tahap pertama

*Scraping* tahap pertama dilakukan untuk artikel berita mulai dari tanggal 1 September 2023 sampai dengan tanggal 10 Januari 2024. *Scraping* dilakukan menggunakan *list link index* berita sesuai dengan rentang waktu yang ditentukan. Hasil *scraping* berupa judul, tanggal, *link* dan meta data setiap berita pada rentang tahun tersebut 10.519 data.

1. *Scraping* tahap kedua

*Scraping* tahap kedua dilakukan untuk artikel berita mulai dari tanggal 11 Januari 2024 sampai dengan tanggal 14 Februari 2024. *Scraping* dilakukan menggunakan hasil *scraping link index* sebelumnya sesuai dengan rentang waktu yang ditentukan. Sama seperti hasil *scraping* tahap pertama, pada tahap kedua hasil *scraping* berupa judul, tanggal, *link* dan meta data setiap berita pada rentang tahun tersebut dengan jumlah total 4.500 data.

1. Penggabungan data

Tahap terakhir merupakan penggabungan hasil *scraping* tahap pertama dan tahap kedua. Sehingga rentang waktu total yaitu mulai dari tanggal 1 September 2023 hingga 14 Februari 2024 dengan jumlah artikel yang didapatkan yaitu berjumlah 15.019 artikel berita. Kemudian hasil *scraping* disimpan dalam bentuk *file* .csv untuk diolah pada tahap *preprocessing* data.

## 3.3 Preprocessing Data

Tahap *preprocessing* data dilakukan untuk memproses data mentah yang didapatkan sebelumnya menjadi data yang siap digunakan untuk tahap implementasi model. Hasil tahap ini berupa data yang bersih dan terstruktur sesuai dengan kebutuhan pada tahap implementasi model. Tahap *preprocessing* data melewati beberapa proses yaitu *case folding*, *cleaning*, *tokenizing* dan *feature* *extraction*.

### ***Case Folding***

Proses *case folding* yang dilakukan pada penelitian ini dimaksudkan untuk mengubah semua karakter yang ada pada data artikel menjadi huruf kecil atau *lower* *case* . Proses ini dilakukan menggunakan data *preprocessing* hasil dari *scraping* data pada tahap sebelumnya. Contoh hasil proses *tokenizing* ditunjukkan pada Tabel 3.1 sebagai berikut.

Tabel 3.1 Contoh hasil case folding

|  |  |
| --- | --- |
| **Sebelum *Case folding*** | **Sesudah *Case folding*** |
| Kapolda Sulawesi Tengah (Sulteng) Irjen Agus Nugroho memimpin pelaksanaan Gelar Operasional tahun 2024. Dia memaparkan 6 sasaran prioritas tahun ini. | kapolda sulawesi tengah (sulteng) irjen agus nugroho memimpin pelaksanaan gelar operasional tahun 2024. dia memaparkan sasaran prioritas tahun ini. |

Berdasarkan contoh hasil proses *case folding* seperti yang ditunjukkan pada Tabel 3.1, semua huruf yang ada pada data sebelum *case folding* diubah menjadi *lower case* atau huruf kecil seperti pada kata “Kapolda” sebelumnya huruf “K” merupakan huruf besar. Kemudian setelah dilakukan *case folding* huruf “k” menjadi huruf kecil dan keseluruhan kata menjadi “kapolda”.

### ***Cleaning***

Proses *cleaning* yang dilakukan pada penelitian ini dimaksudkan untuk menghilangkan karakter selain huruf dan angka yang ada pada kalimat seperti tanda baca [32]. Proses ini dilakukan menggunakan data hasil *case folding* pada proses sebelumnya. Contoh hasil proses *cleaning* ditunjukkan pada Tabel 3.2 sebagai berikut.

Tabel 3.2 Contoh hasil Cleaning

|  |  |
| --- | --- |
| **Sebelum *Cleaning*** | **Sesudah *Cleaning*** |
| kapolda sulawesi tengah (sulteng) irjen agus nugroho memimpin pelaksanaan gelar operasional tahun 2024. dia memaparkan sasaran prioritas tahun ini. | kapolda sulawesi tengah sulteng irjen agus nugroho memimpin pelaksanaan gelar operasional tahun 2024 dia memaparkan sasaran prioritas tahun ini |

Berdasarkan contoh hasil proses *cleaning* seperti yang ditunjukkan pada Tabel 3.2, semua simbol, tanda baca ataupun angka yang bukan termasuk dalam huruf alfabet dihapus. Sehingga menyisakan hanya data yang berisi huruf saja. Seperti pada simbol kurung buka dan tutup pada kata “(sulteng)” menjadi “sulteng” dan tanda baca titik pada kata “2024.” dihapus menjadi “2024”.

### ***Tokenizing***

Proses *tokenizing* yang dilakukan pada penelitian ini dimaksudkan untuk menghilangkan *white* *space* atau spasi pada kalimat dengan mengubah setiap kata ke dalam bentuk daftar token [32][33]. Proses ini dilakukan menggunakan data hasil *cleaning* pada proses sebelumnya. Contoh hasil proses *tokenizing* ditunjukkan pada Tabel 3.3 sebagai berikut.

Tabel 3.3 Contoh hasil tokenizing

|  |  |
| --- | --- |
| **Sebelum *Tokenizing*** | **Sesudah *Tokenizing*** |
| kapolda sulawesi tengah sulteng irjen agus nugroho memimpin pelaksanaan gelar operasional tahun 2024 dia memaparkan sasaran prioritas tahun ini | ['kapolda', 'sulawesi', 'tengah', 'sulteng', 'irjen', 'agus', 'nugroho', 'memimpin', 'pelaksanaan', 'gelar', 'operasional', 'tahun', ‘2024’, 'dia', 'memaparkan', 'sasaran', 'prioritas', 'tahun', 'ini'] |

Berdasarkan contoh hasil proses *tokenizing* seperti yang ditunjukkan pada Tabel 3.3, semua kata yang ada pada data sebelum *tokenizing* diubah menjadi dimasukkan token kata. Kemudian token – token yang dihasilkan dimasukkan ke dalam satu keutuhan *list*.

### ***Stopword Removal***

Proses *stopword removal* yang dilakukan pada penelitian ini dimaksudkan untuk menghilangkan kata – kata umum seperti “yang”, “dan”, dan sebagainya yang tidak penting dalam pemrosesan dan mengurangi dimensi data [32]. Proses ini dilakukan menggunakan data hasil *tokenizing* pada proses sebelumnya. Contoh hasil proses *stopword removal* ditunjukkan pada Tabel 3.4 sebagai berikut.

Tabel 3.4 Contoh hasil stopword removal

|  |  |
| --- | --- |
| **Sebelum S*topword Removal*** | **Sesudah S*topword Removal*** |
| ['kapolda', 'sulawesi', 'tengah', 'sulteng', 'irjen', 'agus', 'nugroho', 'memimpin', 'pelaksanaan', 'gelar', 'operasional', 'tahun', ‘2024’, 'dia', 'memaparkan', 'sasaran', 'prioritas', 'tahun', 'ini'] | ['kapolda', 'sulawesi', 'sulteng', 'irjen', 'agus', 'nugroho', 'memimpin', 'pelaksanaan', 'gelar', 'operasional', ‘2024’, 'memaparkan', 'sasaran', 'prioritas'] |

Berdasarkan contoh hasil proses *stopword removal* seperti yang ditunjukkan pada Tabel 3.4, semua kata umum seperti kata hubung dan kata bantu dihapus dari *list* token yang dihasilkan dari proses *tokenizing* sebelumnya seperti kata ‘ini’, ‘dia’ dan lain sebagainya yang tidak dibutuhkan dalam proses pemodelan topik akan dihapus karena tidak relevan.

## 3.4 Implementasi Model

Dalam proses implementasi model, BERT dipilih sebagai model utama untuk pemodelan topik Pemilu 2024. Keputusan ini didasarkan pada kemampuan BERT dalam menangani tugas pemahaman bahasa alami dan kemampuannya untuk mengekstrak representasi fitur yang berkualitas tinggi dari teks. Pemodelan topik dilakukan menggunakan pustaka BERTopic. BERTopic tidak hanya mampu menghasilkan representasi *embedding* dengan dimensi yang optimal tanpa perlu *dimensionality* *reduction* tambahan, tetapi juga dapat langsung digunakan untuk *document* *clustering* dan *keyword* *extraction*. Dengan menggunakan representasi *embedding* dari BERTopic, kita dapat secara efektif melakukan analisis topik berita dengan mengidentifikasi pola tematik, mengelompokkan artikel berita yang serupa, dan mengekstrak kata kunci penting secara otomatis. Pendekatan ini memungkinkan pendekatan yang ringkas dan efisien tanpa perlu memperkenalkan algoritma tambahan seperti *clustering* atau TF-IDF dikarenakan pada pustaka BERTopic sudah dilakukan secara otomatis menggunakan standar yang telah ditetapkan.

Hasil data sebelum dilakukan pemodelan topik dapat dilihat pada Tabel 3.3 sebelumnya pada kolom sesudah *tokenizing.* Hasil pemodelan topik menggunakan BERTopic menghasilkan jumlah topik yang tidak pasti, karena penggunaan model *clustering* HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) sebagai standar, sehingga jumlah *cluster* yang dihasilkan tidak dapat ditentukan. Model tersebut juga menghasilkan topik “-1” mengacu pada semua *outlier* dan biasanya diabaikan [9]. Sehingga pada penelitian ini akan menggunakan model lainnya yaitu *K-means* agar tidak menghasilkan topik “-1” atau *outlier* karena merupakan model non-hierarki dengan jumlah *cluster* berdasarkan evaluasi *Elbow* [46]. Contoh hasil setelah dilakukan pemodelan topik menggunakan BERTopic ditunjukkan pada Tabel 3.5

Tabel 3.5 Contoh hasil topic modelling BERTopic

|  |  |  |
| --- | --- | --- |
| ***Topic*** | ***Count*** | ***Topic Representation*** |
| 0 | 5181 | 0\_ganjar\_prabowo\_imin\_anies |
| 1 | 2675 | 1\_presiden\_jokowi\_indonesia\_nomor |
| 2 | 2510 | 2\_politik\_survei\_demokrat\_partai |
| 3 | 2348 | 3\_2024\_pemilu\_pilpres\_kpu |
| 4 | 2305 | 4\_debat\_gibran\_nomor\_cawapres |

Tahap implementasi *topic modelling* menggunakan pustaka BERTopic seperti yang ditunjukkan pada Tabel 3.5, hasil pemodelan topik berisi detail dari semua topik yang dibuat oleh BERTopic seperti jumlah atau count yang mewakili jumlah kemunculan topik tertentu yang ada pada dataset dan representasi topik dalam bentuk kata-kata kunci yang mewakili topik tersebut. Sebagai contoh topik “0” memiliki jumlah kemunculan 5181 data dengan kata kunci representasi “ganjar”, “prabowo”, “imin” dan “anies”.

## 3.4 Evaluasi Model

Setelah model BERTopic diimplementasikan, langkah evaluasi dilakukan untuk mengukur sejauh mana model dapat memodelkan topik Pemilu 2024 dengan akurat. Proses ini menggunakan metode *Elbow*, *Silhouette Coefficient, Davies Bouldin Index, Calinski Harabasz Index* dan *Dunn Index* untuk model *clustering* serta metode *topic* *coherence* untuk mengevaluasi hasil pemodelan topik. Semakin besar nilai tersebut maka semakin baik nilai asosiasi hasil pemodelan topik tersebut. Contoh hasil evaluasi pemodelan topik menggunakan BERTopic ditunjukkan pada Tabel 3.6.

Tabel 3.6 Contoh hasil evaluasi topic coherence BERTopic

|  |  |
| --- | --- |
| ***Topic*** | ***Topic Coherence*** |
| 0 | 0.101277 |
| 1 | 0.077433 |
| 2 | 0.065047 |
| 3 | 0.100189 |
| 4 | 0.088987 |

Hasil evaluasi *topic coherence* untuk hasil pemodelan topik menggunakan BERTopic seperti yang ditunjukkan Tabel 3.6, dapat diketahui untuk nilai *topic coherence* pada setiap topik yang dihasilkan. Kemudian dari setiap nilai *topic coherence* dapat dihitung rata – rata untuk mengetahui keseluruhan nilai *topic coherence* dari untuk 5 topik, yaitu menghasilkan nilai 0.0865866.

## 3.6 *Deployment*

Proses terakhir adalah *deployment*, di mana semua proses yang dilakukan pada pemodelan *topic* menggunakan BERTopic dikemas lebih sederhana agar hasil yang didapatkan memberikan data yang informatif dan menarik dalam bentuk sistem dasbor data. Dasbor tersebut dapat digunakan untuk mengidentifikasi topik-topik yang mendominasi perbincangan terkini seputar Pemilu 2024 pada berbagai portal berita *online*. Hasil tersebut akan diimplementasikan pada sistem sederhana dalam bentuk *website* yang memberikan informasi hasil implementasi model dan tren pemberitaan yang mendominasi pada waktu tertentu. Sistem pada penelitian ini dibuat menggunakan *Framework* Streamlit dalam bahasa pemrograman Python.

## 3.7 Alat dan Bahan

Alat dan bahan pada penelitian ini berupa perangkat keras dan perangkat lunak yang digunakan untuk melakukan tugas sesuai dengan kebutuhan pada penelitian ini. Berikut komponen perangkat keras dan perangkat lunak yang digunakan oleh penulis.

1. Perangkat Keras : Processor Intel ® Core ™ i7-8650U CPU, RAM 16 GB  
    dan SSD 512 GB
2. Perangkat Lunak : Google Colaboration, Google Chrome Browser, dan Visual Studio Code

## 3.8 Perancangan Sistem

Perancangan sistem merupakan pembuatan desain *user interface* yang memberikan gambaran tampilan pada sistem yang dirancang pada penelitian ini. Berikut desain *user interface* setiap menu pada sistem.

1. Menu *home*

Menu *home* memberikan informasi mengenai detail topik modeling yang dibuat pada penelitian ini, seperti informasi *dataset* sampai dengan model yang digunakan. Desain *user interface* untuk menu tren pemberitaan ditunjukkan pada Gambar 3.2 sebagai berikut.

A white rectangular box with black text

Description automatically generated

Gambar 3.2 Desain menu tren pemberitaan

Pada menu *home* seperti yang ditunjukkan pada Gambar 3.2, pengguna dapat mengetahui informasi detail terkait proyek yang dilakukan pada penelitian ini. Informasi detail tersebut berupa sumber *dataset* (data hasil *scraping* yang digunakan pada penelitian ini atau data meta deskripsi artikel selama September 2023 hingga 14 Februari 2024 pada portal berita Detik.com), jumlah dataset, model pemodelan topik dan lain sebagainya terkait teknis penelitian ini serta informasi tambahan seperti tujuan penelitian dan lain sebagainya.

1. Menu BERTopic

Menu BERTopic merupakan menu untuk melakukan pemodelan topik menggunakan metode – metode yang telah dijabarkan sebelumnya. Menu tersebut ditunjukkan pada Gambar 3.3 sebagai berikut.

A screenshot of a computer

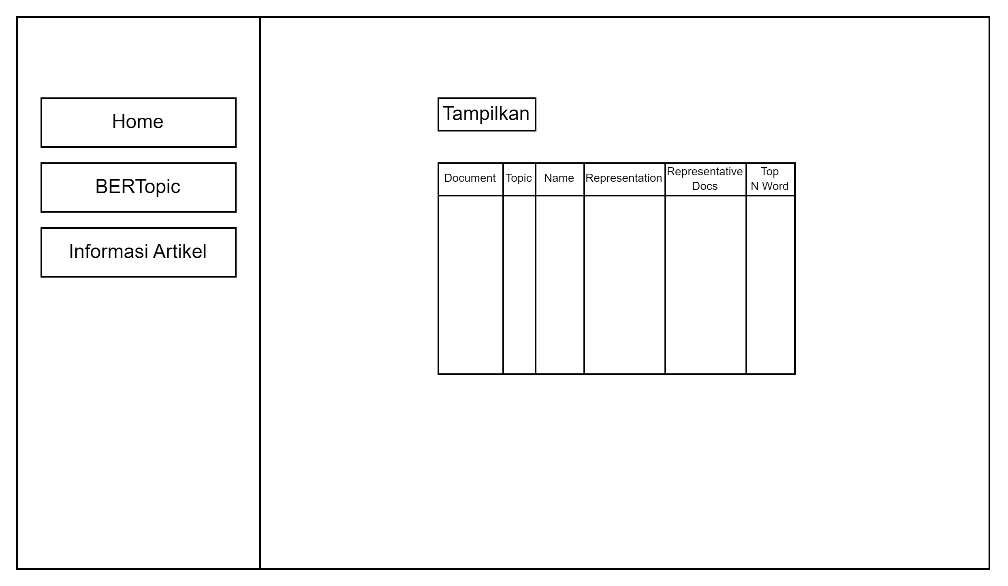
Description automatically generated

Gambar 3.3 Desain menu data mentah

Pada menu BERTopic seperti yang ditunjukkan pada Gambar 3.3, memberikan informasi kepada pengguna terkait tahapan serta hasil pemodelan topik yang dilakukan pada penelitian ini. Meni ini menunjukkan hasil setiap tahapan mulai dari preprocessing data, pembuatan model clustering sampai dengan hasil evaluasi pemodelan topik.

1. Menu informasi artikel

Menu informasi artikel memberikan informasi tentang dokumen hasil pemodelan topik menggunakan BERTopic. Desain *user interface* untuk menu informasi artikelditunjukkan pada Gambar 3.4 sebagai berikut.



Gambar 3.4 Desain menu data preprocessing

Pada menu informasi artikel seperti yang ditunjukkan pada Gambar 3.4, memberikan informasi kepada pengguna terkait hasil pemodelan topik dari setiap data. Hasil yang ditampilkan mulai dari data meta deskripsi atau *document*, *topic* yang berisi jenis topik, *name* yang berisi nama topik, *representation* yang berisi representasi dari setiap jenis topik, *representative docs* yang berisi dokumen perwakilan dari setiap topik dan *top n word* yang berisi kata – kata yang sering muncul pada jenis topik.

# DAFTAR PUSTAKA

[1] Ahmad Arif, “Memperebutkan Gen Z dan Milenial,” Kompas.id.

[2] Christiany Juditha dan Josep J. Darmawan, “PENGGUNAAN MEDIA DIGITAL DAN PARTISIPASI POLITIK  GENERASI MILENIAL,” vol. 22, 2018.

[3] Emilsyah Nur, “PERAN MEDIA MASSA DALAM MENGHADAPI SERBUAN MEDIA ONLINE,” 2021.

[4] C. C. Aggarwal dan C. C. Aggarwal, *Mining Text Data*. Springer, 2015.

[5] Putri Septiani dan Hendra Kurniawan, “Analisa Penggunaan Keyword Untuk Implementasi Search  Engine Optimization (SEO),” vol. XV, 2020.

[6] Roman Egger dan Joanne Yu, “A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts,” 2022.

[7] J. Devlin, M.-W. Chang, K. Lee, dan K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Okt 2018, [Daring]. Tersedia pada: http://arxiv.org/abs/1810.04805

[8] G. Malki, “Efficient Sentiment Analysis and Topic Modeling in NLP using Knowledge Distillation and Transfer Learning.” 2023.

[9] Maarten Grootendorst, “BERTopic.” Diakses: 27 Maret 2024. [Daring]. Tersedia pada: https://maartengr.github.io/BERTopic

[10] Enny Fitriani Nasution, “KOMPARASI *K-MEANS* CLUSTERING DAN LATENT DIRICHLET ALLOCATION PADA KASUS SHORT TEXT TOPIC MODELLING,” 2023.

[11] S. E. Uthirapathy dan D. Sandanam, “Topic Modelling and Opinion Analysis On Climate Change Twitter Data Using LDA And BERT Model.,” *Procedia Comput Sci*, vol. 218, hlm. 908–917, 2023.

[12] A. Abuzayed dan H. Al-Khalifa, “BERT for Arabic topic modeling: An experimental study on BERTopic technique,” *Procedia Comput Sci*, vol. 189, hlm. 191–194, 2021.

[13] I. K. T. Mertayasaa dan I. D. M. B. A. Darmawana, “Pemodelan Topik Pada Ulasan Hotel Menggunakan Metode BERTopic Dengan Prosedur c-TF-IDF,” *Jurnal Nasional Teknologi Informasi dan Aplikasinya* , vol. 1, no. 1, hlm. 307–316, 2022.

[14] E. Puspita, D. F. Shiddieq, dan F. F. Roji, “Pemodelan Topik pada Media Berita Online Menggunakan Latent Dirichlet Allocation (Studi Kasus Merek Somethinc),” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, hlm. 481–489, 2024.

[15] M. Amien, “Sejarah dan Perkembangan Teknik Natural Language Processing (NLP) Bahasa Indonesia: Tinjauan tentang sejarah, perkembangan teknologi, dan aplikasi NLP dalam bahasa Indonesia,” *arXiv preprint arXiv:2304.02746*, 2023.

[16] S. D. Lestari dan A. Z. Falani, “Implementasi Fitur Assistance Virtual Sebagai Media Pembelajaran Interaktif,” *Jurnal RESTIKOM: Riset Teknik Informatika dan Komputer*, vol. 5, no. 1, hlm. 46–54, 2023.

[17] Y. Matira dan I. Setiawan, “Pemodelan Topik pada Judul Berita Online Detikcom Menggunakan Latent Dirichlet Allocation,” *Estimasi: Journal of Statistics and Its Application*, hlm. 53–63, 2023.

[18] I. Vayansky dan S. A. P. Kumar, “A review of topic modeling methods,” *Inf Syst*, vol. 94, hlm. 101582, 2020.

[19] D. Ridhwanullah, “Pemodelan Topik pada Cuitan tentang Penyakit Tropis di Indonesia dengan Metode Latent Dirichlet Allocation,” 2022.

[20] D. Yolanda, M. H. Hersyah, dan E. Marozi, “Implementasi Metode Unsupervised Learning Pada Sistem Keamanan Dengan Optimalisasi Penyimpanan Kamera IP,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 6, hlm. 1099–1105, 2021.

[21] D. P. Sari, F. Sembiring, D. P. Sulisdianto, Y. Jentiner, dan E. Suratno, “Implementasi Algoritma Fuzzy C-Means Untuk Memprediksi Faktor Siswa Membolos (Studi Kasus: Faktor Siswa Membolos Di Smpn 1 Parakansalak),” *Jurnal RESTIKOM: Riset Teknik Informatika dan Komputer*, vol. 2, no. 1, hlm. 1–7, 2020.

[22] N. Wakhidah, “Clustering menggunakan *k-means* algorithm,” *Jurnal Transformatika*, vol. 8, no. 1, hlm. 33–39, 2010.

[23] S. Setyaningtyas, B. I. Nugroho, dan Z. Arif, “Tinjauan Pustaka Sistematis: Penerapan Data Mining Teknik Clustering Algoritma *K-means*,” *Jurnal Teknoif Teknik Informatika Institut Teknologi Padang*, vol. 10, no. 2, hlm. 52–61, 2022.

[24] D. N. Chandra, G. Indrawan, dan I. N. Sukaraja, “Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram,” *Jurnal Ilmiah Teknologi Informasi Asia*, vol. 10, no. 1, hlm. 11–19, 2016.

[25] A. M. Mansah, “Tren Pergeseran Media Konvensional Ke Era Digitalisasi (Studi Kasus Konvergensi Media Di Lembaga Kantor Berita Nasional Antara Biro Sulawesi Selatan-Sulawesi Barat),” *Al-MUNZIR*, vol. 12, no. 1, hlm. 121–130, 2019.

[26] B. Kurniawan, A. A. Aldino, dan A. R. Isnain, “Sentimen Analisis Terhadap Kebijakan Penyelenggara Sistem Elektronik (Pse) Menggunakan Algoritma Bidirectional Encoder Representations From Transformers (Bert),” *J. Teknol. dan Sist. Inf*, vol. 3, no. 4, hlm. 98–106, 2022.

[27] S. Nabilah, “Analisis Sentimen Berbasis Aspek Pada Ulasan Aplikasi Novel Online Di Media Sosial Menggunakan Latent Dirichlet Allocation Dan Bidirectional Encoder Representations From Transformers,” Perpustakaan Fakultas Sains dan Teknologi UIN Jakarta, 2022.

[28] A. Josi dan L. A. Abdillah, “Penerapan teknik web scraping pada mesin pencari artikel ilmiah,” *arXiv preprint arXiv:1410.5777*, 2014.

[29] V. A. Flores, P. A. Permatasari, dan L. Jasa, “Penerapan web scraping sebagai media pencarian dan menyimpan artikel ilmiah secara otomatis berdasarkan keyword,” *Maj. Ilm. Teknol. Elektro*, vol. 19, no. 2, hlm. 157, 2020.

[30] D. Normawati dan S. A. Prayogi, “Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter,” *J-SAKTI (Jurnal Sains Komputer Dan Informatika)*, vol. 5, no. 2, hlm. 697–711, 2021.

[31] J. Supriyanto, D. Alita, dan A. R. Isnain, “Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Analisis Sentimen Publik Terhadap Pembelajaran Daring,” *Jurnal Informatika dan Rekayasa Perangkat Lunak*, vol. 4, no. 1, hlm. 74–80, 2023.

[32] M. Nashrullah dan D. A. Efrilianda, “Sentiment Analysis of Independent Campus Policy on Twitter Using Support Vector Machine and Naïve Bayes Classifier,” *Journal of Advances in Information Systems and Technology*, vol. 4, no. 1, hlm. 13–23, 2022.

[33] A. Sujjada dan A. Fergina, “Implementasi Metode Vector Space Model Untuk Deteksi Emosi Menggunakan Data Teks Twitter,” *Jurnal RESTIKOM: Riset Teknik Informatika dan Komputer*, vol. 3, no. 3, hlm. 116–129, 2021.

[34] Y. An, D. Kim, J. Lee, H. Oh, J.-S. Lee, dan D. Jeong, “Topic Modeling-Based Framework for Extracting Marketing Information From E-Commerce Reviews,” *IEEE Access*, vol. 11, hlm. 135049–135060, 2023.

[35] M. Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure,” *arXiv preprint arXiv:2203.05794*, 2022.

[36] A. Alamsyah dan N. D. Girawan, “Improving clothing product quality and reducing waste based on consumer review using RoBERTa and BERTopic language model,” *Big Data and Cognitive Computing*, vol. 7, no. 4, hlm. 168, 2023.

[37] R. N. Fahmi, M. Jajuli, dan N. Sulistiyowati, “Analisis pemetaan tingkat kriminalitas di kabupaten Karawang menggunakan Algoritma *K-means*,” *INTECOMS: Journal of Information Technology and Computer Science*, vol. 4, no. 1, hlm. 67–79, 2021.

[38] R. B. Trianto, A. S. Nugroho, dan E. Supriyadi, “Klasterisasi Menggunakan Algoritma *K-means* Dan *Elbow* Pada Opini Masyarakat Tentang Kebijakan Sekolah Luring Tahun 2022,” *Jurnal Inovtek Polbeng Seri Informatika*, vol. 8, no. 1, hlm. 1–13, 2023.

[39] M. S. Pangestu dan M. A. Fitriani, “Perbandingan Perhitungan Jarak Euclidean Distance, Manhattan Distance, dan Cosine Similarity dalam Pengelompokan Data Bibit Padi Menggunakan Algoritma *K-means*,” *Sainteks*, vol. 19, no. 2, hlm. 141–155, 2022.

[40] H. Fitriyah, E. M. Safitri, N. Muna, M. Khasanah, D. A. Aprilia, dan D. Nurdiansyah, “IMPLEMENTASI ALGORITMA CLUSTERING DENGAN MODIFIKASI METODE *ELBOW* UNTUK MENDUKUNG STRATEGI PEMERATAAN BANTUAN SOSIAL DI KABUPATEN BOJONEGORO,” *Jurnal Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 4, no. 3, hlm. 1598–1607, 2023.

[41] D. Adimanggala, F. A. Bachtiar, dan E. Setiawan, “Evaluasi Topik Tersembunyi Berdasarkan Aspect Extraction menggunakan Pengembangan Latent Dirichlet Allocation,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 3, hlm. 511–519, 2021.

[42] Made Anditya Mahesastra dan Dewa Made Bayu Atmaja Darmawana, “Pemodelan Topik Teks Berita Menggunakan DistilBERT,” vol. 1, no. 1, 2022.

[43] F. Nuraeni, D. Kurniadi, dan G. F. Dermawan, “Pemetaan Karakteristik Mahasiswa Penerima Kartu Indonesia Pintar Kuliah (KIP-K) menggunakan Algoritma *K-means*++,” *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, vol. 11, no. 3, hlm. 437–443, 2023.

[44] R. Hidayati, A. Zubair, A. H. Pratama, dan L. Indana, “Analisis *Silhouette Coefficient* pada 6 Perhitungan Jarak *K-means* Clustering,” *Techno. Com*, vol. 20, no. 2, hlm. 186–197, 2021.

[45] D. E. Herwindiati dan T. Handhayani, “Clustering Data Covid-19 Di Indonesia Menggunakan Intelligent *K-means*,” *Jurnal Ilmu Komputer dan Sistem Informasi*, vol. 10, no. 2, 2022.

[46] L. Qadrini, “Metode *K-means* dan DBSCAN pada Pengelompokan Data Dasar Kompetensi Laboratorium ITS Tahun 2017,” *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*, vol. 13, no. 2, hlm. 5–11, 2020.

[47] A. P. Riani, A. Voutama, dan T. Ridwan, “Penerapan *K-means* clustering dalam pengelompokan hasil belajar peserta didik dengan metode *Elbow*,” *Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD*, vol. 6, no. 1, hlm. 164–172, 2023.

[48] N. T. Hartanti, “Metode *Elbow* dan *K-means* Guna Mengukur Kesiapan Siswa SMK Dalam Ujian Nasional,” *J. Nas. Teknol. dan Sist. Inf*, vol. 6, no. 2, hlm. 82–89, 2020.

[49] N. Azmi, H. S. Hafsah, Y. Yuyun, dan H. Hazriani, “Penerapan Metode *K-means* Clustering Dalam Mengelompokkan Data Penjualan Obat pada Apotek M23,” *Prosiding SISFOTEK*, vol. 7, no. 1, hlm. 244–248, 2023.

[50] R. Dwirahmanto dan A. Bisri, “Menentukan Nilai K Pada Metode *K-means* Menggunakan Teknik Grid Search Untuk Strategi Produk Pakaian Medis,” *Jurnal Informatika Multi*, vol. 1, no. 2, hlm. 93–103, 2023.

[51] I. T. Umagapi, B. Umaternate, H. Hazriani, dan Y. Yuyun, “Uji Kinerja *K-means* Clustering Menggunakan Davies-Bouldin Index Pada Pengelompokan Data Prestasi Siswa,” *Prosiding SISFOTEK*, vol. 7, no. 1, hlm. 303–308, 2023.

[52] A. M. Sikana dan A. W. Wijayanto, “Analisis Perbandingan Pengelompokan Indeks Pembangunan Manusia Indonesia Tahun 2019 dengan Metode Partitioning dan Hierarchical Clustering,” *J. Ilmu Komput*, vol. 14, no. 2, hlm. 66–78, 2021.

[53] D. A. Saidah, R. Santoso, dan T. Widiharih, “Pengelompokan Provinsi Di Indonesia Berdasarkan Indikator Kesehatan Lingkungan Menggunakan Metode Partitioning Around Medoids Dengan Validasi Indeks Internal,” *Jurnal Gaussian*, vol. 11, no. 2, hlm. 302–312, 2022.

[54] D. R. S. Saputro, “Algoritme Partitioning Around Medoid (Pam) Dengan Calinski-Harabasz Index Untuk Clustering Data Outlier,” *UNEJ e-Proceeding*, hlm. 22–29, 2022.

[55] A. Putranto, N. L. Azizah, dan I. R. I. Astutik, “Sistem Prediksi Penyakit Jantung Berbasis Web Menggunakan Metode SVM dan Framework Streamlit,” *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen)*, vol. 4, no. 2, hlm. 442–452, 2023.

[56] I. Amal, E. W. Pamungkas, S. Kom, dan M. Kom, “Aplikasi Pendeteksi Berita Palsu Bahasa Indonesia Menggunakan Framework Flask dan Streamlit serta Algoritma Machine Learning,” Universitas Muhammadiyah Surakarta, 2023.