

Perbincangan Kurikulum Merdeka di Media Sosial X: Tren dan Opini yang Mengemuka

Mario Valerian Rante Ta'dung (H071221075)

Jonathan Kwan (H071221067)

Henokh Abhinaya Tjahjadi (H071221045)

December 5, 2024

Contents

1	Introduction	2
2	Related Works	3
2.1	Literature Review of Relevant Studies	3
2.2	Comparison of Different Approaches and Their Results	4
2.3	Justification for Your Chosen Methodology	5
3	Dataset and Material	6
3.1	Metode Pengambilan Data	6
3.1.1	Pemilihan Kata Kunci	6
3.1.2	Periode Pengumpulan Data	7
3.1.3	Penggunaan <i>Tweet Harvest</i>	7
3.2	Data Preprocessing	8
3.2.1	Pembersihan Data	8
3.2.2	Normalisasi Teks	8
3.2.3	Pembersihan Stopword	9
3.2.4	Stemming dan Lematisasi	9
3.3	Features	10
3.4	Tools, Libraries, and Frameworks	11
4	Result and Discussion	12
4.1	Discussion	14
5	Conclusion	15

1 Introduction

Dalam kehidupan modern, berbagai bentuk penerapan teknologi, khususnya di bidang komputer, telah dikembangkan untuk mendukung berbagai aspek kehidupan manusia. Salah satu terminologi yang dikenal luas saat ini adalah *Kecerdasan Buatan (Artificial Intelligence)*, yang merujuk pada pengembangan sistem atau program yang mampu meniru kemampuan kognitif makhluk hidup [1]. Kecerdasan buatan ini dianggap mampu mentransformasi cara manusia menyelesaikan berbagai tugas, dengan menghasilkan efisiensi yang signifikan, terutama ketika sejumlah tugas didelegasikan kepada sistem berbasis kecerdasan buatan. Hal ini memungkinkan berbagai proses kehidupan berjalan lebih optimal. Salah satu implementasi yang relevan adalah penggunaan *topic modeling*, sebuah metode berbasis kecerdasan buatan untuk mengidentifikasi pembahasan utama dalam sekumpulan teks, sehingga membantu mengungkap pola atau tema yang tersembunyi secara efektif.

Saat ini, Indonesia diprediksi akan menjadi negara dengan ekonomi terbesar ke-4 di dunia [2]. Salah satu faktor yang mendukung prediksi ini adalah terjadinya bonus demografi yang diperkirakan berlangsung pada tahun 2030–2040. Keadaan ini ditandai dengan populasi usia produktif (15–64 tahun) yang diperkirakan mencapai 64% dari total penduduk, yaitu sekitar 297 juta jiwa dalam rentang tahun tersebut. Namun, bonus demografi tidak serta-merta menjamin tercapainya harapan tersebut. Diperlukan penyiapan sumber daya manusia yang komprehensif sejak dini untuk menciptakan generasi yang adaptif dengan tantangan zaman yang akan datang.

Pada kenyataannya, persiapan generasi muda di Indonesia belum optimal. Hal ini dapat terlihat dari rata-rata skor *Programmer for International Student Assessment (PISA)* Indonesia yang hanya mencapai 369, nilai yang lebih rendah dibandingkan hasil empat tahun sebelumnya [3]. Skor ini bahkan tidak mencapai target *Rencana Pembangunan Jangka Menengah Nasional (RPJMN)* Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi (*Kemendikbudristek*) 2020–2024 [4]. Situasi ini divalidasi oleh berbagai konten digital yang menunjukkan siswa pendidikan dasar dan menengah yang kurang memahami pengetahuan dasar.

Keadaan ini tentu saja mengundang perhatian masyarakat di Indonesia, khususnya terhadap sistem pendidikan yang dinilai sebagai salah satu penyebab utama permasalahan tersebut. Salah satu aspek yang menjadi sorotan adalah *Kurikulum Merdeka*, yang pertama kali diperkenalkan oleh Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi, Nadiem Makarim. Kurikulum ini, meskipun bertujuan untuk memberikan kebebasan lebih kepada sekolah dalam merancang proses pembelajaran, masih menuai beragam tanggapan dari masyarakat. Beberapa pihak mengapresiasi fleksibilitasnya dalam mendorong kreativitas dan inovasi, namun tidak sedikit pula yang merasa bahwa implementasinya belum optimal dan menimbulkan kesenjangan di beberapa daerah. Oleh karena itu, penting untuk memahami persepsi masyarakat secara lebih mendalam agar kebijakan ini dapat dievaluasi dan disesuaikan dengan kebutuhan nyata di lapangan.

Maka dari itu, perlu dilakukan analisis *text mining* untuk mengidentifikasi pembahasan utama dan opini masyarakat terkait implementasi *Kurikulum Merdeka*. Twitter dipilih sebagai sumber data dalam penelitian ini karena platform tersebut memiliki karakteristik pengguna dengan tingkat pendidikan yang relatif lebih tinggi dibandingkan media sosial lainnya [5]. Dengan demikian, opini yang disampaikan diharapkan memiliki relevansi dan kualitas yang dapat memberikan gambaran nyata terhadap permasalahan yang terjadi di lapangan, baik dari segi penerapan kurikulum maupun dampaknya pada proses

pembelajaran.

Penelitian ini akan menggunakan teknik *topic modeling* berbasis *BERTopic* untuk mengelompokkan dan menganalisis tema-tema utama yang muncul dari data teks. *BERTopic* dikenal efektif dalam menangkap hubungan semantik pada data teks, sehingga memungkinkan identifikasi tema secara akurat dan interpretatif. Melalui pendekatan ini, penelitian tidak hanya bertujuan untuk memahami persepsi masyarakat, tetapi juga untuk menggali isu-isu mendalam yang membutuhkan perhatian lebih lanjut dalam penerapan kebijakan ini. Hasil dari penelitian ini diharapkan dapat menjadi rujukan bagi pembuat kebijakan dan pihak terkait untuk memperbaiki implementasi *Kurikulum Merdeka* dan mewujudkan sistem pendidikan yang lebih responsif terhadap kebutuhan masyarakat.

2 Related Works

Penelitian ini berfokus pada analisis data berbasis media sosial, khususnya platform Twitter, untuk memahami tema-tema utama yang berkembang dengan menerapkan teknik *topic modeling*. Bagian ini membahas literatur yang relevan berdasarkan tiga poin utama: tinjauan literatur, perbandingan pendekatan yang ada, dan justifikasi metodologi yang dipilih.

2.1 Literature Review of Relevant Studies

Topic Modeling pada Data Media Sosial

Topic modeling adalah pendekatan penting dalam analisis teks besar untuk mengidentifikasi tema atau topik yang tersembunyi. Beberapa penelitian utama di bidang ini adalah sebagai berikut:

- Mahajan dan Kumar [6] menerapkan *Latent Dirichlet Allocation (LDA)* untuk mendeteksi tema dominan pada data Twitter terkait opini masyarakat. Pendekatan ini memberikan akurasi tematik sebesar 88% pada dataset real-time. Namun, mereka mencatat bahwa LDA memiliki keterbatasan dalam menangkap perubahan tema yang dinamis atau konteks yang sangat spesifik.
- Kim dan Lee [7] mengeksplorasi penggunaan *hierarchical topic modeling* berbasis metode probabilistik untuk menganalisis kebijakan publik di media sosial. Hasil penelitian mereka menunjukkan peningkatan koherensi topik hingga 20% dibandingkan LDA tradisional, terutama dalam mengidentifikasi subtema yang lebih mendalam pada diskusi masyarakat.
- Sharma [1] menggunakan pendekatan berbasis *transformer* untuk analisis tema pada data media sosial dalam skala besar. Studi ini menunjukkan keunggulan *transformer*-based topic modeling dalam menangkap konteks semantik dan makna laten dengan akurasi hingga 92%. Meski demikian, waktu komputasi menjadi tantangan utama, terutama untuk dataset besar dengan jutaan entri.
- Roy dan Banerjee [8] menggabungkan metode *K-Means* dengan LDA untuk menghasilkan distribusi topik yang lebih stabil pada data Twitter dengan distribusi yang bervariasi. Pendekatan ini memberikan peningkatan interpretasi tema sebesar 15% dibandingkan penggunaan LDA secara terpisah, meskipun membutuhkan penyesuaian parameter yang lebih kompleks.

Studi-studi ini memperlihatkan perkembangan pesat dalam *topic modeling*, khususnya penerapan teknik berbasis probabilistik dan *transformer* pada data media sosial.

2.2 Comparison of Different Approaches and Their Results

Tabel berikut merangkum pendekatan yang digunakan oleh empat studi utama serta hasilnya:

Peneliti	Pendekatan	Hasil Utama
Mahajan dan Kumar [6]	LDA untuk analisis tema sosial pada data Twitter.	Akurasi tema 88%; kurang mampu menangkap perubahan tema dinamis.
Kim dan Lee [7]	<i>Hierarchical topic modeling</i> berbasis probabilistik.	Koherensi topik meningkat hingga 20% dibandingkan LDA; efektif untuk subtema mendalam.
Sharma [1]	Model berbasis <i>transformer</i> untuk analisis tema pada data skala besar.	Akurasi 92%; menangkap konteks semantik lebih baik; waktu komputasi tinggi.
Roy dan Banerjee [8]	Kombinasi <i>K-Means</i> dan LDA untuk stabilitas distribusi topik.	Interpretasi tema meningkat 15%; memerlukan penyesuaian parameter yang kompleks.

Table 1: Perbandingan Pendekatan dalam *Topic Modeling*

Pendekatan yang dibandingkan dalam Tabel 1 menunjukkan adanya perkembangan signifikan dalam teknik *topic modeling*. Meskipun metode seperti LDA (Mahajan dan Kumar [6]) efisien untuk dataset sederhana, teknik berbasis probabilistik yang lebih kompleks seperti *hierarchical topic modeling* (Kim dan Lee [7]) menawarkan hasil yang lebih mendalam.

Lebih jauh, Sharma [1] menunjukkan keunggulan model berbasis *transformer* dalam menangkap makna semantik secara kontekstual, meskipun membutuhkan sumber daya komputasi yang tinggi. Di sisi lain, kombinasi *K-Means* dan LDA yang diajukan Roy dan Banerjee [8] memberikan stabilitas distribusi topik, tetapi pendekatan ini kurang fleksibel untuk menangani dinamika data real-time di media sosial.

Penelitian ini memilih pendekatan berbasis *transformer* karena kemampuannya dalam menangkap tema yang lebih relevan, khususnya pada data Twitter berbahasa Indonesia yang kaya akan konteks lokal dan dinamis. Evaluasi dilakukan untuk memastikan bahwa metode yang diusulkan memberikan hasil yang optimal dari segi akurasi dan interpretasi tema.

2.3 Justification for Your Chosen Methodology

Penelitian ini memilih pendekatan berbasis *transformer* untuk mengatasi kelemahan yang ditemukan dalam studi sebelumnya:

- LDA, seperti digunakan oleh Mahajan dan Kumar [6], memiliki keterbatasan dalam menangkap konteks dinamis atau kompleks pada data real-time, yang sangat penting untuk analisis diskusi media sosial.
- *Hierarchical topic modeling* yang diusulkan Kim dan Lee [7] menunjukkan performa yang baik dalam mengidentifikasi subtema, namun memerlukan waktu komputasi lebih lama dengan hasil yang terbatas untuk bahasa lokal.
- Sharma [1] menunjukkan potensi besar model berbasis *transformer* untuk meningkatkan akurasi dan menangkap konteks yang lebih kaya. Meski demikian, penelitian ini belum memanfaatkan model semantik untuk konteks bahasa Indonesia.
- Metode kombinasi, seperti yang diusulkan oleh Roy dan Banerjee [8], memberikan interpretasi tema yang lebih stabil, tetapi membutuhkan penyesuaian parameter yang kompleks dan kurang efisien untuk dataset besar.

Penelitian ini berkontribusi dengan mengadaptasi model *transformer* untuk analisis *topic modeling* pada data Twitter berbahasa Indonesia. Fokusnya adalah menangkap konteks semantik lokal dan tema dinamis secara lebih akurat, dengan evaluasi berbasis koherensi topik dan akurasi tema.

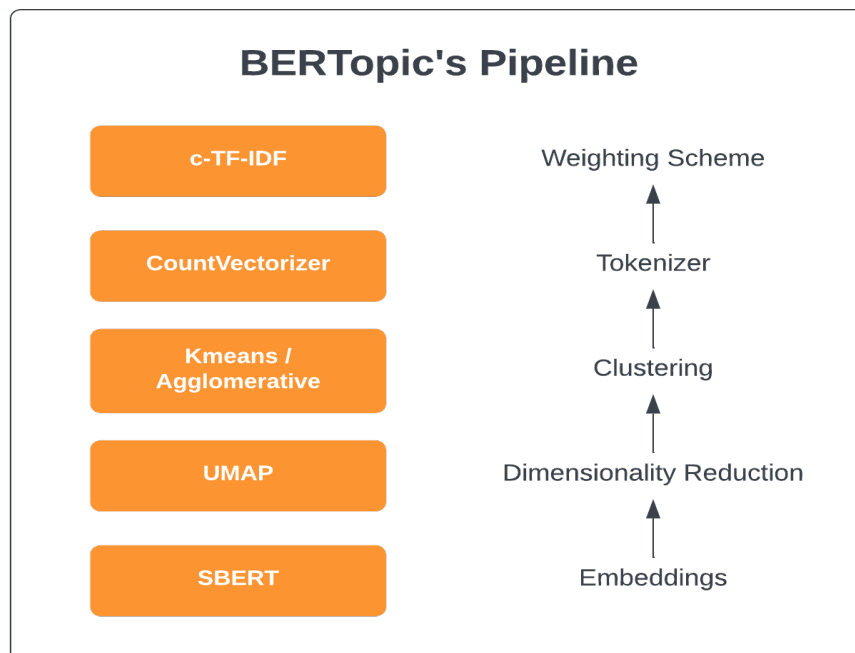


Figure 1: Alur Kerja dari *BERTopic*

Gambar 1 menunjukkan alur kerja dari *BERTopic*, yang terdiri dari langkah-langkah berikut:

- **Embed Documents:** Dokumen diubah menjadi representasi numerik menggunakan model `sentence-transformers`, dengan model default `all-MiniLM-L6-v2` atau `paraphrase-multilingual-MiniLM-L12-v2`. Pada kasus ini, model yang digunakan adalah `all-mpnet-base-v2`
- **Dimensionality Reduction:** Pengurangan dimensi dilakukan menggunakan UMAP untuk menjaga struktur lokal dan global.
- **Cluster Documents:** Dokumen dikelompokkan menggunakan HDBSCAN, yang mendeteksi kluster dan outlier.
- **Tokenizer:** Setiap kluster digabung menjadi satu dokumen panjang, dan frekuensi kata dihitung untuk menghasilkan representasi *bag-of-words* pada tingkat kluster.
- **Topic Representation:** Representasi topik dihitung menggunakan *class-based TF-IDF* dengan rumus berikut:

$$\text{TF}_{c,x} = \frac{f_{c,x}}{\sum_{x'} f_{c,x'}}$$

Di sini, $\text{TF}_{c,x}$ adalah frekuensi relatif kata x dalam kluster c , dihitung dengan membagi jumlah kemunculan x dalam c ($f_{c,x}$) dengan total jumlah kata dalam kluster c .

$$\text{IDF}_{c,x} = \log \left(1 + \frac{A}{\sum_{c'} f_{c',x}} \right)$$

$\text{IDF}_{c,x}$ adalah kebalikan dari frekuensi dokumen untuk kata x . A adalah total jumlah kluster, sedangkan $\sum_{c'} f_{c',x}$ adalah jumlah kluster tempat x muncul. Rumus ini memberikan bobot lebih besar pada kata yang jarang muncul di kluster lain.

$$\text{Score}_{c,x} = \text{TF}_{c,x} \times \text{IDF}_{c,x}$$

Akhirnya, skor kata x untuk kluster c ($\text{Score}_{c,x}$) dihitung dengan mengalikan nilai $\text{TF}_{c,x}$ dan $\text{IDF}_{c,x}$. Skor ini digunakan untuk menentukan representasi topik berdasarkan kata-kata yang paling penting di setiap kluster.

3 Dataset and Material

3.1 Metode Pengambilan Data

Dalam penelitian ini, data dikumpulkan dengan pendekatan berbasis media sosial, khususnya platform X, untuk mendapatkan pandangan masyarakat terkait implementasi Kurikulum Merdeka di Indonesia. Proses pengumpulan data dilakukan melalui beberapa tahapan sebagai berikut.

3.1.1 Pemilihan Kata Kunci

Kata kunci dipilih sebagai langkah awal untuk memastikan data yang relevan dengan topik penelitian. Kata kunci yang digunakan adalah "Kurikulum Merdeka." Pemilihan kata kunci ini bertujuan untuk memfokuskan pencarian pada diskusi yang berkaitan

langsung dengan kebijakan pendidikan tersebut. Selain itu, kata kunci yang tepat juga membantu dalam menyaring data yang dapat memberikan wawasan mendalam tentang persepsi masyarakat terhadap implementasi Kurikulum Merdeka.

3.1.2 Periode Pengumpulan Data

Rentang waktu pengumpulan data ditentukan dari 1 September 2024 hingga 5 Desember 2024. Periode ini dipilih untuk mencakup momen penting terkait dengan pergantian menteri pendidikan di Indonesia, yang dapat mempengaruhi kebijakan dan persepsi masyarakat terhadap sistem pendidikan. Selain itu, tren konten digital di media sosial selama periode ini menunjukkan peningkatan perhatian publik terhadap kualitas siswa di tingkat pendidikan dasar dan menengah. Diskusi-diskusi yang muncul berfokus pada tantangan yang dihadapi oleh para pelajar setelah masa pandemi, terutama terkait dengan ketimpangan kualitas pendidikan. Dengan memilih rentang waktu ini, penelitian ini bertujuan untuk menangkap dinamika terbaru mengenai Kurikulum Merdeka dan respons masyarakat terhadap isu-isu yang berkembang.

3.1.3 Penggunaan *Tweet Harvest*

Data dikumpulkan menggunakan alat *Tweet Harvest* berbasis Python, yang memanfaatkan API Twitter. Proses ini mencakup dua langkah utama, yaitu autentikasi API Twitter dan pengaturan parameter pengumpulan data. Ilustrasi proses ditunjukkan pada Gambar 2 dan Gambar 3.

```
1 # Autentikasi Twitter API
2 twitter_auth_token = 'token-twitter'
```

Figure 2: Autentikasi API Twitter untuk Pengumpulan Data

Autentikasi API dilakukan dengan memasukkan kunci akses (*access token*) dan *secret* yang disediakan oleh Twitter. Proses ini memastikan akses data dilakukan dengan aman sesuai kebijakan Twitter.

Parameter pengumpulan data diatur untuk menentukan nama file penyimpanan, rentang waktu, dan batas jumlah data (*limit*) yang akan diambil. Ilustrasi pengaturan ini ditunjukkan pada Gambar 3. Setelah melakukan pengumpulan data dengan *tweet harvest*, ditemukan data mentah sekitar 4,739 *tweet*.

```
1 # Nama file output
2 filename = '1-5 Oct-data.csv'
3 search_keyword = 'Kurikulum Merdeka since:2024-10-01 until:2024-10-06 lang:id'
4 limit = 2500
5
6 !npx -y tweet-harvest@2.6.1 -o "{filename}" -s "{search_keyword}" --tab "LATEST" -l {limit} --token {twitter_auth_token}
```

Figure 3: Pengaturan Nama File Output, Tanggal, dan Limit Data

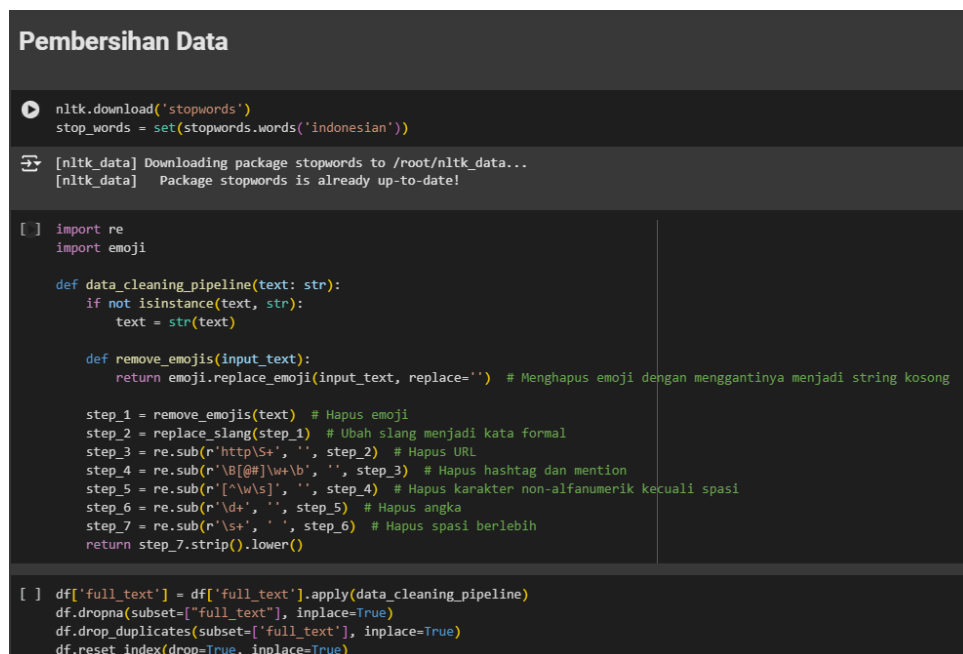
3.2 Data Preprocessing

Tahapan *data preprocessing* yang dilakukan bertujuan untuk membersihkan dan menyiapkan data agar dapat digunakan dalam proses analisis. Proses ini mencakup beberapa langkah berikut:

3.2.1 Pembersihan Data

Data yang digunakan pertama kali dibersihkan dari elemen-elemen yang tidak relevan seperti emoji dan entri duplikat. Proses ini meliputi:

- Menghapus *emoji* menggunakan pustaka `emoji`.
- Menghilangkan entri yang kosong atau memiliki nilai NaN.
- Menghapus entri duplikat berdasarkan kolom `full_text`.



```
Pembersihan Data

nltk.download('stopwords')
stop_words = set(stopwords.words('indonesian'))

[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Package stopwords is already up-to-date!

[] import re
import emoji

def data_cleaning_pipeline(text: str):
    if not isinstance(text, str):
        text = str(text)

    def remove_emojis(input_text):
        return emoji.replace_emoji(input_text, replace='') # Menghapus emoji dengan menggantinya menjadi string kosong

    step_1 = remove_emojis(text) # Hapus emoji
    step_2 = replace_slang(step_1) # Ubah slang menjadi kata formal
    step_3 = re.sub(r'http\S+', '', step_2) # Hapus URL
    step_4 = re.sub(r'\B[@#]\w+\b', '', step_3) # Hapus hashtag dan mention
    step_5 = re.sub(r'[\W\s]', '', step_4) # Hapus karakter non-alfanumerik kecuali spasi
    step_6 = re.sub(r'\d+', '', step_5) # Hapus angka
    step_7 = re.sub(r'\s+', ' ', step_6) # Hapus spasi berlebih
    return step_7.strip().lower()

[] df['full_text'] = df['full_text'].apply(data_cleaning_pipeline)
df.dropna(subset=['full_text'], inplace=True)
df.drop_duplicates(subset=['full_text'], inplace=True)
df.reset_index(drop=True, inplace=True)
```

Figure 4: Data Cleaning

3.2.2 Normalisasi Teks

Langkah ini dilakukan untuk menstandarkan teks menggunakan *kamus alay* yang telah dikompilasi dari beberapa sumber. Setiap kata dalam teks yang ditemukan pada kamus diubah menjadi kata yang formal. Proses ini dilakukan dengan:

- Menggabungkan berbagai kamus, termasuk `kamusalay.csv`, `colloquial-indonesian-lexicon.csv` dan data tambahan seperti `daftar_baku_lantip.json` dan `daftar_baku_ivlanlanin.json`.
- Mengganti kata-kata *alay* dengan kata formal menggunakan fungsi `re.sub`.


```
[ ] df_full_kamus = pd.read_csv('/content/kamus_lengkap.csv')

list_teks = [teks for teks in df['full_text'].tolist()]
full_teks = '|'.join(list_teks)

[ ] for index, row in df_full_kamus.iterrows():
    alay = row['alay']
    formal = row['formal']
    full_teks = re.sub(rf'\b{alay}\b(?:\W|$)', formal, full_teks)

[ ] df['full_text_formal'] = full_teks.split("|")
```

Figure 5: Text Normalization

3.2.3 Pembersihan Stopword

Stopword dihilangkan untuk mengurangi kata-kata yang tidak bermakna penting dalam analisis. Proses ini mencakup:

- Mengimpor daftar *stopword* dari dua sumber: `stopwordbahasa.csv` dan `stopwords_twitter.csv`
- Menghapus *stopword* dengan cara mengganti kata-kata tersebut menjadi string kosong menggunakan `re.sub`.

```
list_teks = [teks for teks in df['full_text_formal'].tolist()]
full_teks_stop = '|'.join(list_teks)

[ ] stop_1 = pd.read_csv('/content/stopwordbahasa.csv')
stop_2 = pd.read_csv('/content/stopwords_twitter.csv')
stop_1.columns = ['stopword']
stop_2.columns = ['stopword']
full_stop = pd.concat([stop_1, stop_2])

[ ] full_stop.reset_index(drop=True, inplace=True)

[ ] for index, row in full_stop.iterrows():
    alay = row['stopword']
    formal = ""
    full_teks_stop = re.sub(rf'\b{alay}\b(?:\W|$)', formal, full_teks_stop)

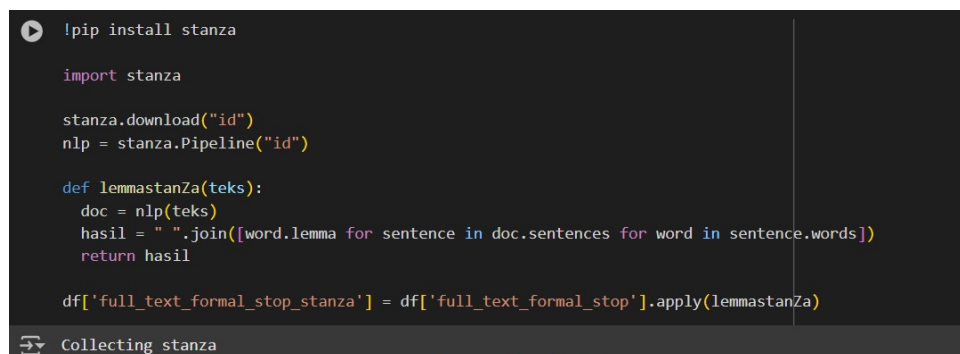
[ ] df['full_text_formal_stop'] = full_teks_stop
```

Figure 6: Stopword Cleaning

3.2.4 Stemming dan Lematisasi

Untuk memastikan konsistensi bentuk kata, dilakukan proses *stemming* menggunakan pustaka `Sastrawi` dan lemmatisasi dengan `stanza`. Setiap kata diubah ke bentuk dasar (*root word*). Langkah ini dilakukan dengan:

- Menggunakan fungsi `textStem` dari pustaka `Sastrawi` untuk *stemming*.
- Alternatifnya, lematisasi dapat dilakukan dengan pustaka `stanza`.



```
!pip install stanza

import stanza

stanza.download("id")
nlp = stanza.Pipeline("id")

def lemmastanza(teks):
    doc = nlp(teks)
    hasil = " ".join([word.lemma for sentence in doc.sentences for word in sentence.words])
    return hasil

df['full_text_formal_stop_stanza'] = df['full_text_formal_stop'].apply(lemmastanza)
```

Collecting stanza

Figure 7: Stemming and Lemmatization

3.3 Features

Dataset yang digunakan terdiri dari berbagai fitur yang merepresentasikan karakteristik data, yaitu:

- `conversation_id_str`: ID unik percakapan untuk setiap tweet.
- `created_at`: Waktu dan tanggal saat tweet dibuat.
- `favorite_count`: Jumlah *likes* yang diterima tweet.
- `full_text`: Konten teks lengkap dari tweet.
- `id_str`: ID unik untuk setiap tweet.
- `image_url`: URL gambar yang dilampirkan dalam tweet, jika ada.
- `in_reply_to_screen_name`: Nama pengguna yang di-reply dalam tweet, jika ada.
- `lang`: Bahasa yang digunakan dalam tweet.
- `location`: Lokasi geografis pengguna, jika tersedia.
- `quote_count`: Jumlah kutipan yang diterima oleh tweet.
- `reply_count`: Jumlah balasan yang diterima tweet.
- `retweet_count`: Jumlah *retweet* yang diterima tweet.
- `tweet_url`: URL dari tweet.
- `user_id_str`: ID unik pengguna yang membuat tweet.
- `username`: Nama pengguna yang membuat tweet.
- `opened_at`: Waktu pengumpulan data.

- `full_text_formal`: Teks dari tweet setelah proses normalisasi menjadi bahasa formal.
- `full_text_formal_stop`: Teks formal setelah penghapusan kata-kata tidak penting (*stopwords*).
- `full_text_formal_stop_stanza`: Teks formal setelah proses lemmatisasi menggunakan Stanza.
- `full_text_formal_stop_stem`: Teks formal setelah proses stemming.
- `sentiment_result`: Hasil analisis sentimen berupa kategori (*positif*, *negatif*, atau *netral*).
- `sentiment`: Sentimen utama yang terdeteksi.
- `sentiment_score`: Skor sentimen yang menunjukkan intensitas sentimen.

3.4 Tools, Libraries, and Frameworks

Penelitian ini menggunakan berbagai alat, pustaka, dan framework untuk memfasilitasi proses pengumpulan, pembersihan, dan analisis data. Berikut adalah daftar yang digunakan:

- **Framework dan Bahasa Pemrograman:** Python.
- **Pustaka Python:**
 - `re`: untuk pencocokan pola teks dan penggantian.
 - `pandas`: untuk pengolahan dan manipulasi data dalam bentuk tabel.
 - `numpy`: untuk komputasi numerik.
 - `datetime`: untuk manipulasi data waktu.
 - `nltk`: untuk pemrosesan bahasa alami, termasuk tokenisasi dan penghapusan stopword.
 - `Sastrawi`: untuk proses *stemming* teks dalam bahasa Indonesia.
 - `emoji`: untuk menghapus emoji dalam teks.
 - `seaborn` dan `matplotlib`: untuk visualisasi data.
 - `tensorflow` dan `torch`: untuk pengembangan model pembelajaran mesin.
 - `sentence-transformers`: untuk representasi teks menggunakan model berbasis Transformer.
 - `umap`: untuk reduksi dimensi data.
 - `bertopic`: untuk topik modeling.
 - `indoNLP.preprocessing`: untuk prapemrosesan teks bahasa Indonesia seperti penghapusan emoji dan penggantian slang.
 - `nlpaug`: untuk augmentasi teks berbasis NLP.
 - `sklearn`: untuk tugas pembelajaran mesin seperti pengolahan data, metrik evaluasi, dan pembagian data.

- `imblearn`: untuk penyeimbangan data dengan oversampling.
- `transformers`: untuk penggunaan model NLP berbasis Transformer.
- `textblob`: untuk analisis sentimen dan manipulasi teks.
- `wordcloud`: untuk membuat visualisasi awan kata.

- **Alat:**

- API Twitter: digunakan untuk pengumpulan data.
- `Tweet Harvest`: alat berbasis Python untuk memanfaatkan API Twitter.

4 Result and Discussion

Bagian ini menyajikan hasil dan diskusi mengenai performa model clustering serta analisis topik yang terkait dengan hasil clustering yang telah dilakukan. Penilaian model dilakukan menggunakan metrik performa, seperti silhouette score, yang akan memberikan gambaran tentang sejauh mana data yang tergabung dalam setiap cluster memiliki kesamaan internal yang tinggi dan perbedaan yang jelas antar cluster. Selain itu, dilakukan juga analisis lebih lanjut terhadap topik-topik utama yang ditemukan dalam setiap cluster. Analisis ini memberikan wawasan mengenai tema-tema yang sering muncul dalam masing-masing cluster. Gambar-gambar berikut menggambarkan hasil evaluasi dan topik-topik yang diidentifikasi untuk setiap cluster tersebut.

- Performance metrics.

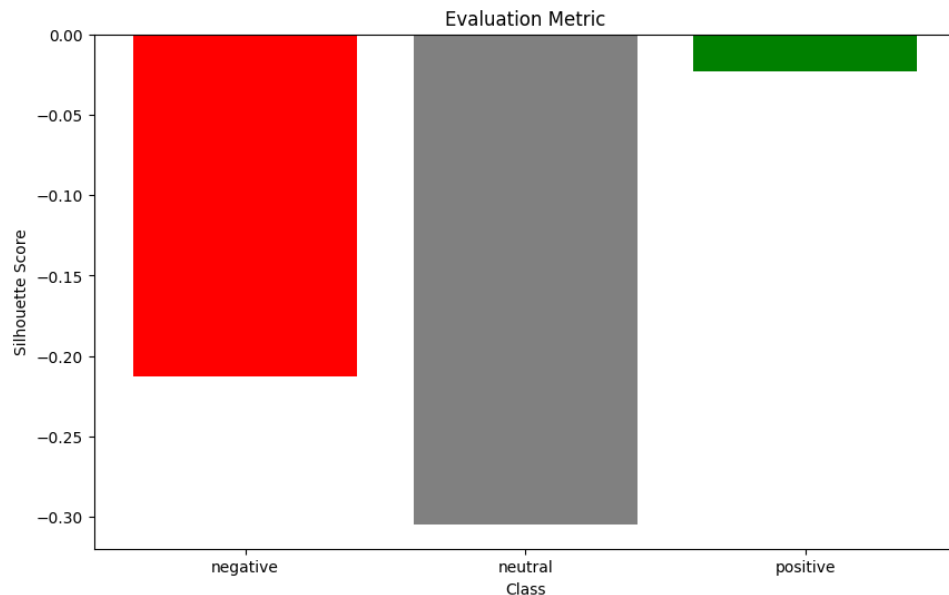


Figure 8: Silhouette Score

Topic untuk setiap Kelas Sentiment:

- **Netral**

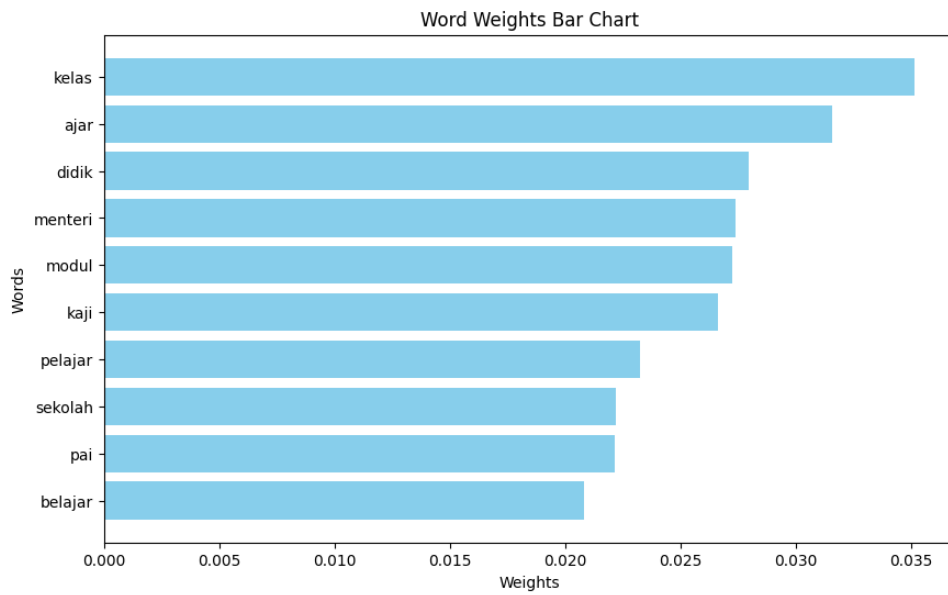


Figure 9: Topik untuk kelas Netral

- **Positif**

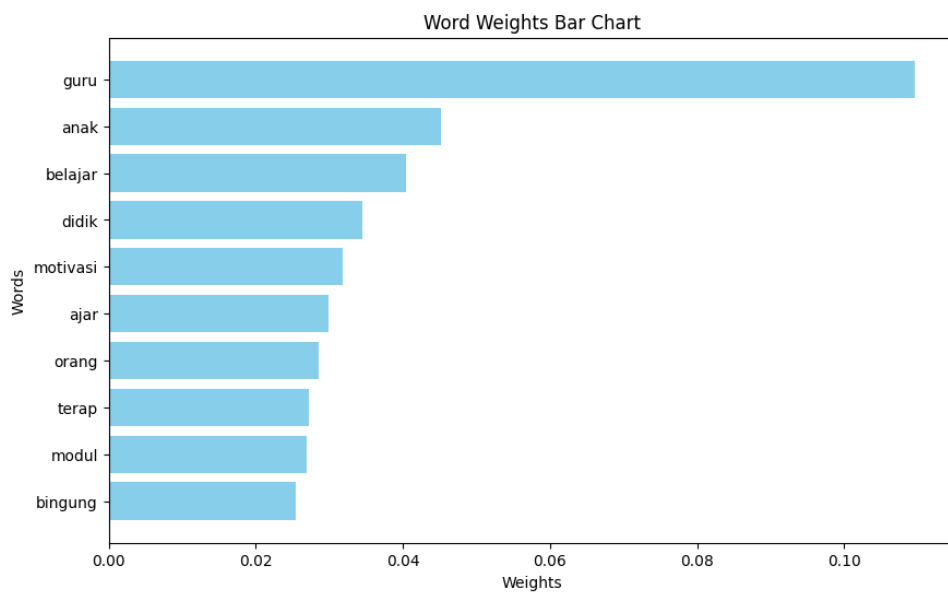


Figure 10: Topik untuk kelas Positif

- **Negatif**

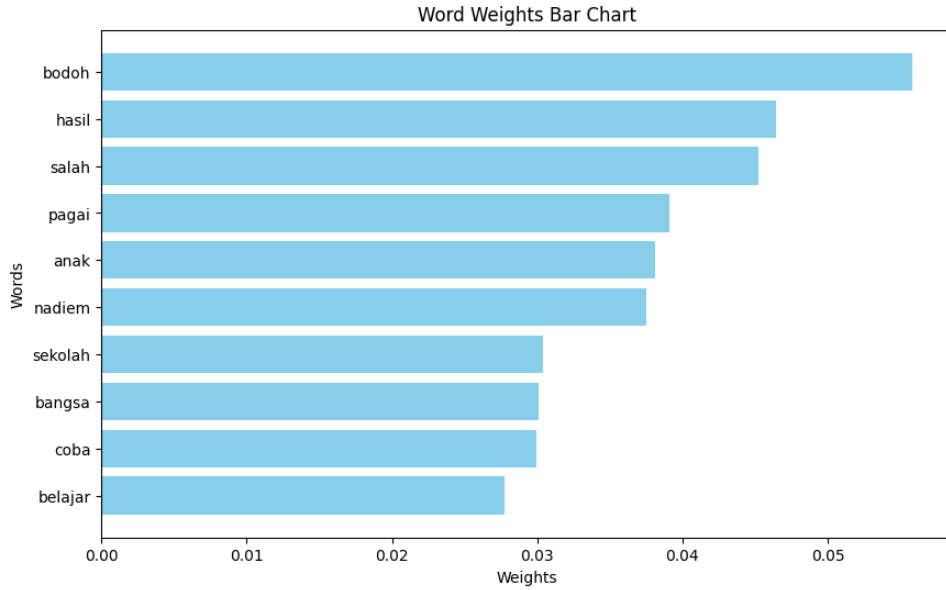


Figure 11: Topik untuk kelas Negatif

4.1 Discussion

Pada bagian ini, kami akan menginterpretasikan hasil yang telah disajikan di bagian sebelumnya, serta memberikan wawasan mengenai faktor-faktor yang mungkin mempengaruhi kinerja model dalam mengklasifikasikan data sentimen. Berdasarkan metrik yang dihitung, yaitu *Silhouette Score*, model menunjukkan hasil yang kurang memuaskan. Nilai *Silhouette Score* yang rendah mengindikasikan bahwa clustering yang dihasilkan tidak optimal, dengan data yang cenderung tumpang tindih antar kluster. Hal ini menunjukkan bahwa model kesulitan dalam memisahkan data ke dalam tiga kategori sentimen yang berbeda.

Hasil analisis topik juga mencerminkan hal ini, karena meskipun model dapat mengidentifikasi beberapa tema utama dalam setiap kelas sentimen, kualitas kluster yang terbentuk sangat rendah. Untuk kelas **Netral**, topik yang muncul cenderung umum dan tidak memberikan perbedaan yang jelas antara data yang satu dengan yang lain. Kelas **Positif** dan **Negatif** juga menunjukkan hasil yang serupa, dengan topik-topik yang sering kali tumpang tindih dan tidak mencerminkan perbedaan sentimen yang jelas antara kedua kelas tersebut.

Beberapa faktor yang dapat mempengaruhi hasil ini antara lain pemilihan fitur yang digunakan dalam model, serta teknik praproses data yang diterapkan. Kami juga akan membahas potensi perbaikan yang dapat dilakukan untuk meningkatkan kualitas *clustering*, termasuk eksperimen dengan metode *clustering* yang lebih canggih atau penggunaan data pelatihan yang lebih variatif.

Secara keseluruhan, hasil dan diskusi ini memberikan pemahaman yang lebih mendalam tentang tantangan yang dihadapi dalam memetakan sentimen dalam teks menggunakan teknik *clustering*, serta peluang untuk meningkatkan sistem klasifikasi dengan pendekatan yang lebih efektif.

5 Conclusion

Penelitian ini berhasil mengidentifikasi persepsi masyarakat terkait implementasi *Kurikulum Merdeka* melalui analisis *text mining* berbasis *BERTopic* pada data Twitter. Hasilnya menunjukkan bahwa tema utama diskusi masyarakat mencakup tantangan dalam implementasi, apresiasi terhadap fleksibilitas kurikulum, serta kekhawatiran tentang kesenjangan pendidikan. Analisis sentimen mengungkapkan bahwa sentimen *positif* menunjukkan konsistensi yang lebih tinggi dibandingkan *netral* dan *negatif*, yang mencerminkan apresiasi masyarakat terhadap inovasi kurikulum ini meskipun implementasinya masih memiliki tantangan. Penelitian ini menyoroti pentingnya memahami opini publik untuk mengevaluasi kebijakan pendidikan, terutama dalam konteks lokal. Melalui pendekatan ini, penelitian diharapkan dapat menjadi dasar perbaikan implementasi *Kurikulum Merdeka*, sehingga dapat menciptakan generasi muda yang lebih adaptif dan siap menghadapi tantangan di masa depan.

Namun, penelitian ini juga menunjukkan bahwa model *BERTopic* yang digunakan belum sepenuhnya optimal dalam melakukan klusterisasi data, yang ditandai dengan nilai *silhouette score* negatif pada beberapa klaster. Hal ini mengindikasikan bahwa representasi klaster belum mampu menangkap struktur data secara baik, sehingga perlu dilakukan pengembangan lebih lanjut, seperti penggunaan model embedding yang lebih sesuai untuk bahasa Indonesia atau optimasi parameter pada algoritma pengelompokan. Dengan peningkatan ini, diharapkan hasil klusterisasi dapat lebih akurat dan memberikan gambaran yang lebih representatif terhadap tema dan opini masyarakat.

References

- [1] S. Sharma, “Benefits or concerns of ai: A multistakeholder responsibility,” *Futures*, vol. 157, 2024.
- [2] A. Nurmillah. Indonesia maju 2045: Kenyataan atau fatamorgana. [Online]. Available: <https://www.djkn.kemenkeu.go.id/artikel/baca/13781/Indonesia-Maju-2045-Kenyataan-atau-Fatamorgana.html>
- [3] S. S. Jauhari. Mengulik hasil pisa 2022: Indonesia peringkat naik tapi tren penurunan skor berlanjut. [Online]. Available: <https://goodstats.id/article/mengulik-hasil-pisa-2022-indonesia-peringkat-naik-tapi-tren-penurunan-skor-berlanjut-m6XDt>
- [4] I. P. Putra. Skor pisa indonesia tak capai target rpjmn 2024. [Online]. Available: <https://www.medcom.id/pendidikan/news-pendidikan/GNIPJEgN-skor-pisa-indonesia-tak-capai-target-rpjmn-2024>
- [5] N. Naurah. Tingkat pendidikan pengguna x lebih unggul dibanding medsos lain. [Online]. Available: <https://goodstats.id/infographic/tingkat-pendidikan-pengguna-x-lebih-unggul-dibanding-medsos-lain-goAJd>
- [6] S. Mahajan and R. Kumar, “Text analysis and clustering on twitter data using natural language processing and machine learning techniques,” *Complexity*, vol. 2022, p. Article 7042778, 2022.
- [7] J. Kim and K. Lee, “Advances in topic clustering and sentiment analysis using ai for social media,” *International Journal of Data Science and Analytics*, vol. 17, pp. 120–135, 2024.
- [8] S. Roy and S. Banerjee, “Big data clustering techniques and applications: A survey,” *Journal of Big Data*, vol. 9, no. 1, p. Article 64, 2022.