JURNAL ILMU KOMPUTER
Volume 4 Nomor 1, Januari, 2026
ISSN: 3031-125X (ONLINE)
Penerbit: Program Studi Teknik Informatika S-2 Universitas Pamulang

Analisis BERT dan LDA Untuk Ekstraksi Kebijakan Ekonomi Presiden Prabowo Subianto

*Muhammad Najwah[1], Sajarwo Anggai[2],Sudarno[3]

[1,2,3)] Teknik Informatika S-2, Program Pascasarjana, Universitas Pamulang, Kota Tangerang Selatan, Banten

*Email: [1]*muhammadnajwah@gmail.com,[2]*sajarwo@gmail.com, [3]*sudarnowiharjo@gmail.com*

ABSTRACT

Economic policies introduced at the beginning of President Prabowo Subianto's administration have generated diverse public discourses reflected in online media coverage. The large volume of textual data necessitates computational approaches to extract information systematically. This study aims to identify, label, and compare major economic policy topics using topic modeling techniques, namely Latent Dirichlet Allocation (LDA) and BERTopic.The dataset consists of 1,000 economic news articles collected through web scraping from an online news portal. Text preprocessing includes normalization, case folding, cleaning, tokenization, and lemmatization. LDA was implemented using a TF-IDF representation and evaluated with the Coherence Score (c_v). BERTopic employed IndoBERT embeddings, UMAP for dimensionality reduction, and HDBSCAN for hierarchical clustering, with evaluation based on topic coherence and semantic interpretability. The results show that LDA generated eight main topics with a Coherence Score (c_v) of 0.61, indicating moderate performance but limited semantic representation, leading to overlapping topics. In contrast, BERTopic produced nine main topics with a higher Coherence Score (c_v) of 0.72 and clearer, more contextual topic labels, including fiscal policy, energy, capital markets, and economic stimulus. Overall, BERTopic outperformed LDA in extracting and labeling economic policy topics due to its superior ability to capture semantic context and form stable topic clusters.

Keywords: Topic Modeling, LDA, BERTopic, Economic Policy, BERT, Topic Extraction.

ABSTRAK

Kebijakan ekonomi yang diterapkan pada awal pemerintahan Presiden Prabowo Subianto memunculkan beragam wacana publik yang tercermin dalam pemberitaan media daring. Besarnya volume data teks tersebut menuntut penggunaan metode komputasional yang mampu mengekstraksi informasi secara sistematis dan bermakna. Penelitian ini bertujuan untuk mengidentifikasi, memberi label, dan membandingkan topik utama kebijakan ekonomi berdasarkan data berita ekonomi menggunakan pendekatan pemodelan topik, yaitu Latent Dirichlet Allocation (LDA) dan BERTopic.. Dataset yang digunakan terdiri dari 1.000 artikel berita ekonomi yang dikumpulkan melalui web scraping dari portal berita daring. Tahapan prapemrosesan teks meliputi normalisasi, case folding, pembersihan data, tokenisasi, dan lemmatisasi. Model LDA dikembangkan menggunakan representasi TF-IDF dan dievaluasi dengan Coherence Score (c_v). Sementara itu, BERTopic memanfaatkan embedding IndoBERT, reduksi dimensi berbasis UMAP, serta klasterisasi hierarkis menggunakan HDBSCAN, dengan evaluasi berdasarkan koherensi topik dan interpretabilitas semantik. Hasil penelitian menunjukkan bahwa LDA menghasilkan delapan topik utama dengan nilai Coherence Score (c_v) sebesar 0,61, yang menunjukkan kinerja cukup baik namun masih terbatas dalam menangkap konteks semantik sehingga terjadi tumpang tindih kosakata antar topik. Sebaliknya, BERTopic menghasilkan sembilan topik utama dengan nilai Coherence Score (c_v) sebesar 0,72 serta label topik yang lebih jelas, kontekstual, dan relevan, seperti kebijakan fiskal, energi, pasar modal, dan stimulus ekonomi. Secara keseluruhan, BERTopic terbukti lebih unggul dibandingkan LDA dalam ekstraksi dan pelabelan topik kebijakan ekonomi.
.

Kata Kunci: Topic Modeling, LDA, BERTopic, Kebijakan Ekonomi, BERT, Ekstraksi Topik.

1. PENDAHULUAN

Sejak masa transisi menuju era pemerintahan Presiden Prabowo Subianto, Indonesia memasuki fase awal transformasi kebijakan ekonomi yang intensif. Fokus utama kebijakan ekonomi nasional mencakup stabilitas fiskal, peningkatan nilai tambah produk dalam negeri melalui hilirisasi industri, penguatan ketahanan pangan dan energi, serta reformasi struktur ekonomi untuk meningkatkan daya saing global. Reformasi kebijakan tersebut ditujukan untuk mengurangi defisit neraca perdagangan, memperkuat basis industri domestik, serta mendorong investasi berkelanjutan di sektor-sektor strategis seperti manufaktur, energi terbarukan, dan teknologi informasi. Kebijakan-kebijakan ini memiliki implikasi besar pada dinamika pasar dan kesejahteraan masyarakat, sehingga setiap langkah kebijakan secara intensif dibahas tidak hanya dalam forum pemerintah tetapi juga melalui ruang publik digital yang mencerminkan persepsi, kritik, dan dukungan masyarakat luas[1].

Peran media menjadi ruang utama dalam pembentukan diskursus publik yang merefleksikan respons masyarakat terhadap kebijakan pemerintah. Portal berita forum diskusi digital menyediakan data tekstual yang besar dan heterogen terkait berita, opini, dan tanggapan publik terhadap kebijakan ekonomi yang diumumkan pemerintah. Data teks ini, dalam jumlah besar dan tak terstruktur, menyimpan pola diskursif yang dapat diungkapkan melalui teknik topic modeling suatu bentuk analisis data teks untuk mengekstraksi tema atau “topik” tersembunyi yang dominan dalam korpus teks yang luas[2].

Era pemerintahan Presiden Prabowo Subianto menandai fase awal transformasi kebijakan ekonomi nasional yang berfokus pada stabilitas fiskal, hilirisasi industri, serta penguatan sektor strategis. Setiap kebijakan yang dirumuskan tidak hanya berdampak pada aspek ekonomi makro, tetapi juga memicu respons publik yang luas melalui media daring. Portal berita menjadi ruang utama terbentuknya diskursus publik yang mencerminkan persepsi, kritik, dan dukungan masyarakat terhadap kebijakan pemerintah[3].

Untuk menjawab tantangan tersebut, penelitian ini menerapkan dan membandingkan dua metode ekstraksi topik terkemuka *Latent Dirichlet Allocation* (LDA) dan *Bidirectional Encoder Representations from Transformers* (BERT)[4]. LDA adalah

algoritma statistik yang telah lama digunakan untuk mengidentifikasi klaster kata yang membentuk topik dalam korpus teks. Meskipun efektif, kelemahan utama LDA adalah pendekatannya yang bersifat probabilistik dan seringkali mengabaikan konteks semantik antar kata, sehingga relevansi topik yang dihasilkan terkadang kurang optimal[5].

Kemajuan pesat dalam *Natural Language Processing* (NLP) dan pembelajaran mesin telah melahirkan metode pemodelan topik berbasis representasi kontekstual, salah satunya BERTopic yang memanfaatkan embedding dari model bahasa berbasis BERT. Berbeda dengan pendekatan probabilistik tradisional, BERTopic menggunakan representasi *vektor* berdimensi tinggi dari teks yang dihasilkan oleh model transformer untuk menangkap nuansa semantik dan hubungan kontekstual antar kata dalam sebuah dokumen. Proses ini biasanya dilanjutkan dengan teknik *clustering* berdasarkan embedding dokumen untuk menghasilkan topik yang lebih semantik dan relevan secara konteks. Metode ini terbukti mampu menangani variasi bahasa informal dan struktur teks yang beragam dalam korpus media daring secara lebih efektif dibandingkan model klasik seperti LDA[6].

Sebagai solusi atas keterbatasan LDA, penelitian ini juga memanfaatkan BERT, sebuah model bahasa modern berbasis transformer yang unggul dalam memahami konteks kalimat secara mendalam[7]. Dengan kemampuannya menangkap nuansa makna, BERT berpotensi menghasilkan pengelompokan topik yang lebih koheren dan relevan secara semantik. Penelitian ini tidak hanya menggunakan kedua model secara terpisah, tetapi juga menganalisis performa keduanya secara komparatif untuk tugas spesifik ekstraksi topik kebijakan ekonomi era Presiden Prabowo Subianto[8].

2. METODE

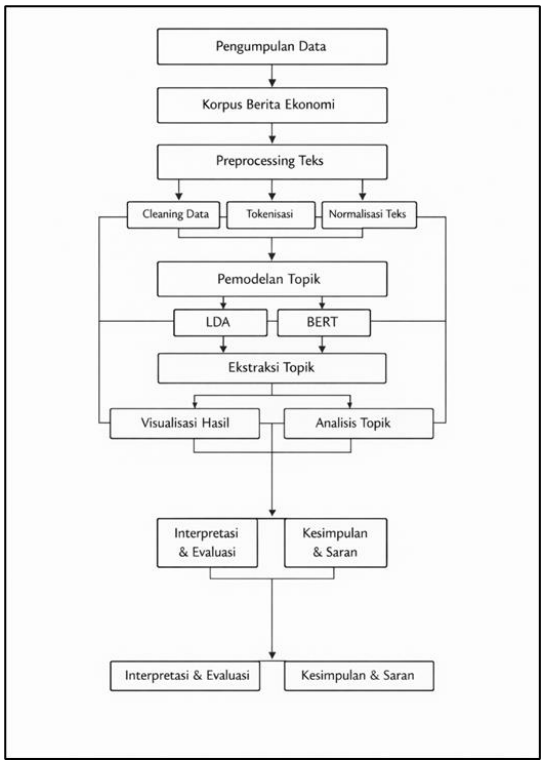
Data penelitian berupa 1.000 artikel berita ekonomi yang diperoleh dari portal berita. Pengumpulan data dilakukan menggunakan teknik *web scraping* dengan kata kunci yang berkaitan dengan kebijakan ekonomi nasional.

Tahapan penelitian diawali dengan proses *preprocessing* teks yang bertujuan untuk membersihkan dan menormalkan data agar siap digunakan dalam pemodelan. Tahapan *preprocessing* meliputi normalisasi teks, *case folding* untuk menyeragamkan huruf, penghapusan karakter *non-alfabet*, tokenisasi, serta lemmatisasi untuk mengembalikan

kata ke bentuk dasarnya. Proses ini penting untuk mengurangi *noise* dan meningkatkan kualitas hasil pemodelan topik.

Pemodelan topik dilakukan menggunakan dua metode, yaitu LDA dan BERTopic. Model LDA dibangun dengan menggunakan representasi TF-IDF dan dilakukan eksperimen untuk menentukan jumlah topik optimal. Evaluasi model LDA dilakukan menggunakan *Coherence Score* (c_v). Sementara itu, BERTopic menggunakan *embedding* IndoBERT untuk merepresentasikan dokumen dalam bentuk vektor, kemudian dilakukan reduksi dimensi menggunakan UMAP dan *clustering* menggunakan HDBSCAN. Evaluasi BERTopic difokuskan pada nilai koherensi dan interpretabilitas topik yang dihasilkan.

Perancangan penelitian bertujuan untuk mendefinisikan langkah-langkah yang digunakan dalam pelaksanaan penelitian agar sesuai dengan tujuan yang telah ditetapkan. Adapun tahap-tahap utama dalam perancangan penelitian seperti Gambar 1 dibawah ini :

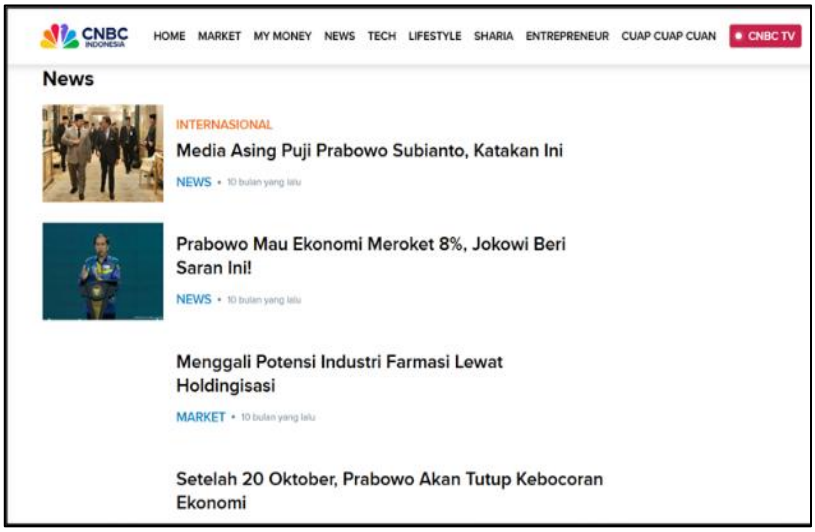


Gambar 1. Kerangka Pemikiran

3. PEMBAHASAN

3.1. Pengambilan Dataset

Data yang dikumpulkan untuk penelitian ini berasal dari sejumlah portal berita online terkemuka di Indonesia yang membahas berbagai masalah kebijakan ekonomi. Tujuan pengumpulan data adalah untuk mendapatkan korpus teks berupa artikel berita yang relevan dengan kebijakan ekonomi pemerintah Indonesia selama periode tertentu. Pengumpulan data dilakukan secara otomatis melalui proses *web crawling* menggunakan bahasa pemrograman Python dengan bantuan pustaka *newspaper3k* dan *BeautifulSoup*. Sebanyak 1000 artikel yang berhasil dikumpulkan dari berbagai sumber berita dan mencakup kebijakan fiskal, moneter, kebijakan anggaran, subsidi, investasi, dan peraturan ekonomi dari awal mulai menjabat, seperta pada tampilan portal pada Gambar 2.



Gambar 2. Portal Berita Ekonomi

3.2. *Scraping* Data dan Penyimpanan Data

Pengumpulan data dalam penelitian ini dilakukan melalui metode *web scraping*, yaitu teknik pengambilan data secara otomatis dari situs web menggunakan bahasa pemrograman Python. Tujuan utama dari *scraping* ini adalah memperoleh artikel-artikel berita dari portal daring nasional yang memuat informasi mengenai kebijakan ekonomi pemerintah Indonesia. Portal berita yang digunakan dalam penelitian ini antara lain Kompas.com, Detik.com, CNN Indonesia.com, dan Liputan6.com, yang

dipilih karena memiliki reputasi tinggi dan cakupan pemberitaan yang luas di bidang ekonomi dan kebijakan publik.

Teknis pelaksanaan scraping dilakukan dengan menggunakan beberapa pustaka Python, yaitu *requests* untuk mengunduh halaman *web, BeautifulSoup* untuk menavigasi dan memproses struktur HTML, serta *newspaper3k* untuk mengekstraksi konten inti dari setiap artikel seperti judul dan isi teks. Proses dimulai dengan mengakses halaman kategori ekonomi pada masing-masing portal, kemudian seluruh tautan yang terdapat di halaman tersebut diidentifikasi. Dari sekian banyak tautan, dilakukan seleksi berdasarkan pola tertentu yang umum dimiliki oleh artikel berita, seperti tautan yang mengandung kata "berita", "*read*", "*news*", atau tahun tertentu "2025", seperti pada Gambar 3.

Gambar 3. Scrip Scrapping Data Berita

3.3. Penggunaan *Library*

Pustaka yang saling terintegrasi digunakan untuk mendukung berbagai tahapan analisis teks, mulai dari *preprocessing* hingga pemodelan topik. Tahapan pra-pemrosesan teks melibatkan penggunaan pustaka seperti NLTK, *spaCy*, dan *Gensim*. NLTK digunakan untuk tokenisasi, penghapusan stopwords, dan lemmatisasi, sedangkan gensim digunakan untuk membangun korpus dan kamus, serta untuk mengevaluasi koherensi model. Untuk pemodelan topik, digunakan *library* BERTopic, yang mendukung pendekatan kontemporer dalam *topic modeling* berbasis

transformer dan *clustering*. Semua integrasi antar *library* dilakukan secara terstruktur untuk memastikan bahwa proses dari input teks hingga hasil visualisasi topik dapat dilakukan dengan akurat. Berikut ditampilkan struktur *library* pada Gambar 4.

```
!pip install pandas
!pip install gensim
!pip install spacy
!pip install scikit-learn
!pip install sentence-transformers
!pip install umap-learn
!pip install hdbscan
!pip install bertopic
!pip install matplotlib
!pip BeautifulSoup
from bs4 import BeautifulSoup
!pip istall re
import re
!pip install numpy
!pip install nltk
import nltk
print("Downloading NLTK data...")
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')
print("Download complete.")
```

Gambar 4. Library Yang digunakan

3.4. *Prepocesing* Data

Pada tahap *preposesing* data bertujuan untuk mensetrukturkan, merapikan,dan memastikan data siap dianalsis dari judul-judul jurnal Teknik Informatika. *Preprocessing* data pada penelitian ini dilakukan tahap Normalization. Berikut *script preprocessing* data dapat dilihat pada Gambar 5.

```
import pandas as pd
import re
from bs4 import BeautifulSoup
text = pd.read_csv("C:\PYTHON\DB_Ekonomi\db_Ekonomii.csv")
df['Content'] = df['Content'].astype(str)
df['Content'] = df['Content'].apply(lambda x: BeautifulSoup(x, 'html.parser').get_text())
def cleaning_text(text):
    text = str(text)                                    # pastikan string
    text = text.lower()                                 # case folding
    text = re.sub(r"http\S+", " ", text)                # hapus URL
    text = re.sub(r"[^a-zA-Z0-9\s]", " ", text)         # hapus simbol
    text = re.sub(r"\s+", " ", text).strip()            # hapus spasi berlebih
    return text
hapus_kata = ["jakarta cnbc indonesia", "liputan6", "cnn indonesia", "kompas"]
df['clean_text'] = df['Content'].apply(cleaning_text)
for kata in hapus_kata:
    df['clean_text'] = df['clean_text'].str.replace(rf"\b{kata}\b", "", regex=True)
# Rapikan lagi spasi berlebih
df['clean_text'] = df['clean_text'].str.replace(r"\s+", " ", regex=True).str.strip()
df.to_csv("hasil_cleaning.csv", index=False, encoding="utf-8-sig")
hapus_kalimat = [
    "jakarta cnbc indonesia",
]
for kalimat in hapus_kalimat:
    df['clean_text'] = df['clean_text'].str.replace(kalimat, "", regex=False)
# Rapikan spasi lagi
df['clean_text'] = df['clean_text'].str.replace(r"\s+", " ", regex=True).str.strip()

print("✅ Hasil cleaning sudah disimpan di 'hasil_cleaning.csv'")
```

Gambar 5. *Script Prepocesing*

3.5. Case Folding

Tahap ini, semua huruf dirubah menjadi huruf kecil. Tahapan *case folding* secara umum adalah dimulai dengan memeriksa ukuran setiap karakter dari awal sampai

akhir karakter, kemudian jika ditemukan karakter yang menggunakan huruf kapital atau *uppercase*, maka huruf tersebut akan diubah menjadi huruf kecil atau *lowercase*, seperti pada Gambar 6.

```
import pandas as pd
import re
from bs4 import BeautifulSoup
text = pd.read_csv("C:\PYTHON\DB_Ekonomi\db_Ekonomii.csv")
df['Content'] = df['Content'].astype(str)
df['Content'] = df['Content'].apply(lambda x: BeautifulSoup(x, 'html.parser').get_text())
def cleaning_text(text):
    text = str(text)                                # pastikan string
    text = text.lower()                             # case folding
    text = re.sub(r"http\S+", " ", text)            # hapus URL
    text = re.sub(r"[^a-zA-Z0-9\s]", " ", text)     # hapus simbol
    text = re.sub(r"\s+", " ", text).strip()        # hapus spasi berlebih
    return text
hapus_kata = ["jakarta cnbc indonesia", "liputan6", "cnn indonesia", "kompas"]
df['clean_text'] = df['Content'].apply(cleaning_text)
for kata in hapus_kata:
    df['clean_text'] = df['clean_text'].str.replace(rf"\b{kata}\b", "", regex=True)
# Rapikan lagi spasi berlebih
df['clean_text'] = df['clean_text'].str.replace(r"\s+", " ", regex=True).str.strip()
df.to_csv("hasil_cleaning.csv", index=False, encoding="utf-8-sig")
hapus_kalimat = [
    "jakarta cnbc indonesia",
]
for kalimat in hapus_kalimat:
    df['clean_text'] = df['clean_text'].str.replace(kalimat, "", regex=False)
# Rapikan spasi lagi
df['clean_text'] = df['clean_text'].str.replace(r"\s+", " ", regex=True).str.strip()

print("✅ Hasil cleaning sudah disimpan di 'hasil_cleaning.csv'")
```

Gambar 6. *Case Folding*

3.6. Cleaning

Proses *cleaning* untuk mengurangi atau membersihkan *corpus* dari kata atau kalimat yang tidak diperlukan seperti tanda baca, *unicode*, dan lain-lain. Proses *cleaning* ini terdapat 5 tahapan yang akan dilakukan oleh sistem untuk memperoleh hasil yang maksimal, seperti membersihkan tanda baca, membersihkan angka, membersihkan link ,hashtag, membersihkan kelebihan spasi dan *URL*, seperti program pada Gambar 7.

```
def cleanin class str(object: object = '')
    text = str(text)                                # pastikan string
    text = text.lower()                             # case folding
    text = re.sub(r"http\S+", " ", text)            # hapus URL
    text = re.sub(r"[^a-zA-Z0-9\s]", " ", text)     # hapus simbol
    text = re.sub(r"\s+", " ", text).strip()        # hapus spasi berlebih
    return text
hapus_kata = ["jakarta cnbc indonesia", "liputan6", "cnn indonesia", "kompas"]
df['clean_text'] = df['Content'].apply(cleaning_text)
for kata in hapus_kata:
    df['clean_text'] = df['clean_text'].str.replace(rf"\b{kata}\b", "", regex=True)
# Rapikan lagi spasi berlebih
df['clean_text'] = df['clean_text'].str.replace(r"\s+", " ", regex=True).str.strip()
```

Gambar 7. *Cleaning*

4. HASIL PEMODELAN TOPIK LDA

4.1. Jumlah Topik

LDA yang dikembangkan dalam penelitian ini menghasilkan sejumlah topik yang merepresentasikan pola hubungan antarkata dalam korpus data secara statistik. Setiap topik tersusun atas kata-kata dengan bobot probabilitas tertentu, yang menunjukkan tingkat kontribusi kata tersebut terhadap pembentukan makna topik. Melalui perintah *print_topics*, model menampilkan sepuluh kata dengan bobot tertinggi pada setiap topik, sehingga mempermudah proses interpretasi konseptual. Analisis terhadap kata-kata kunci ini memungkinkan peneliti mengidentifikasi tema dominan, keterkaitan semantik, serta struktur tematik yang muncul dari keseluruhan data.

Hasil pemetaan topik ini memberikan dasar analitis yang kuat bagi proses kategorisasi dan penarikan kesimpulan berbasis data. Visualisasi distribusi topik dan struktur kata pada setiap topik ditampilkan pada Gambar 8.

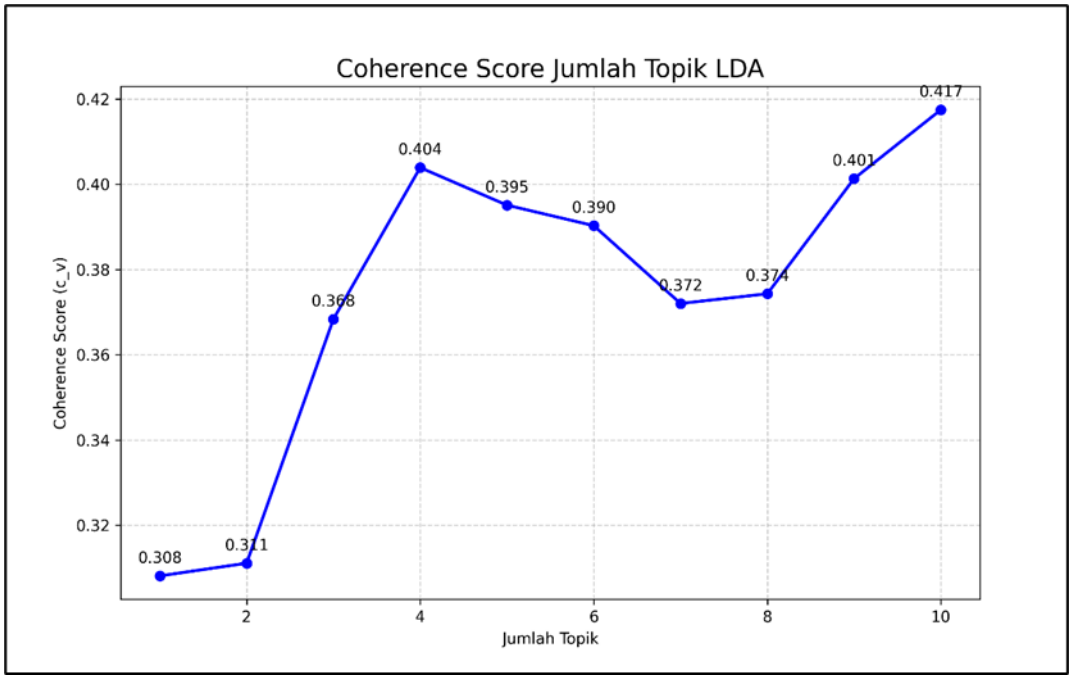
```
===== HASIL TOPIK LDA =====

Topik 0 : 0.002*"riding" + 0.002*"expansion" + 0.002*"wave" + 0.002*"kupas" + 0.002*"indonesiaeconomic" + 0.002*"tu
Topik 1 : 0.002*"tkdn" + 0.001*"meutya" + 0.001*"samsung" + 0.001*"iki" + 0.001*"relokasi" + 0.001*"odol" + 0.001*"
Topik 2 : 0.001*"bos" + 0.001*"nawardi" + 0.001*"bhr" + 0.001*"purn" + 0.001*"nasihat" + 0.001*"ajib" + 0.001*"wira
Topik 3 : 0.004*"diskon" + 0.004*"upah" + 0.003*"digital" + 0.003*"buruh" + 0.003*"brics" + 0.003*"subsidi" + 0.002
Topik 4 : 0.002*"hgbt" + 0.001*"mmbtu" + 0.001*"gakkum" + 0.001*"hekal" + 0.001*"peti" + 0.001*"hanif" + 0.001*"adi
Topik 5 : 0.001*"ptbi" + 0.001*"siddhi" + 0.001*"shoigu" + 0.000*"baro" + 0.000*"permata" + 0.000*"klm" + 0.000*"sa
Topik 6 : 0.003*"tarif" + 0.003*"saham" + 0.003*"trump" + 0.003*"dagang" + 0.002*"pasar" + 0.002*"tumbuh" + 0.002*"
Topik 7 : 0.001*"kemenperin" + 0.001*"nomura" + 0.001*"jemmy" + 0.001*"sumitro" + 0.001*"skenario" + 0.001*"ruptl" +
Topik 8 : 0.001*"kuota" + 0.001*"sritex" + 0.001*"purnomo" + 0.001*"bmi" + 0.001*"bpd" + 0.001*"roy" + 0.001*"irs" +
Topik 9 : 0.001*"ricky" + 0.001*"jkp" + 0.001*"edi" + 0.001*"isy" + 0.001*"ksti" + 0.001*"kppu" + 0.001*"piter" + 0.
```

Gambar 8. Hasil Topik

4.2. Hasil *Coherence*

Coherence score dihitung untuk menentukan jumlah topik yang paling optimal dalam model LDA. Pengujian dilakukan untuk jumlah topik 1 hingga 10. Nilai *coherence* mengalami kenaikan bertahap dari topik 1 (0.308) hingga mencapai titik puncak pada topik 10 dengan nilai 0.417, yang merupakan nilai paling tinggi dalam keseluruhan rentang pengujian. Kenaikan nilai *coherence* menunjukkan bahwa semakin besar jumlah topik, model mampu memisahkan struktur semantik dokumen dengan lebih baik, hingga mencapai topik ke-10 di mana model menghasilkan kelompok kata paling koheren.

Setelah titik itu, jumlah topik tidak diuji lebih lanjut karena sudah terlihat kecenderungan naik yang stabil. Visualisasi nilai pada setiap topik ditampilkan pada Gambar 9.



Gambar 9. *Coherence Score*

4.3. TF-IDF

Perhitungan bobot *Term Frequency–Inverse Document Frequency* (TF-IDF) pada korpus penelitian ini menghasilkan representasi numerik yang menunjukkan tingkat pentingnya suatu kata dalam sebuah dokumen relatif terhadap keseluruhan korpus. Bobot TF-IDF yang tinggi menunjukkan bahwa kata tersebut memiliki frekuensi kemunculan yang kuat pada dokumen tertentu namun tidak muncul secara merata di seluruh dokumen, sehingga kata tersebut dianggap lebih informatif dalam membedakan konteks atau tema dokumen. Sebaliknya, kata-kata dengan bobot rendah merupakan kata yang muncul secara umum di banyak dokumen sehingga kontribusi informatifnya relatif kecil**,** seperti pada tabel 1.

Tabel 1. Hasil TF-IDF

No	Dokumen	Kata	TF-IDF
1	0	ihsg	0.557955
2	0	saham	0.315348
3	0	tbk	0.278604
4	0	terjun	0.194195
5	0	konglomerat	0.175173
6	0	level	0.168330
7	0	euforia	0.141739

8	0	pesta	0.136977
9	0	lesat	0.123997
10	0	dagang	0.122493

5. HASIL PEMODELAN TOPIK DENGAN BERTopic

5.1. Embedding

Pemodelan *embedding* pada penelitian ini menggunakan *SentenceTransformer* dengan arsitektur *paraphrase-multilingual-MiniLM-L12-v2*, yang diinisialisasi melalui perintah *embedding_model = SentenceTransformer("paraphrase-multilingual-MiniLM-L12-v2")*. Model ini merupakan bagian dari keluarga transformer-based sentence *embedding* yang dirancang untuk memetakan teks ke dalam ruang vektor berdimensi tetap, sehingga setiap kalimat atau dokumen direpresentasikan sebagai vektor numerik yang merefleksikan makna semantik secara kontekstual. Pendekatan ini melampaui metode representasi berbasis frekuensi kata, karena *embedding* yang dihasilkan mempertimbangkan urutan kata, konteks, serta hubungan semantik antar token dalam suatu teks**.** Selain itu, karakteristik multilingual pada model ini menjadikannya sangat relevan untuk penelitian yang menggunakan data berbahasa Indonesia, karena model telah dilatih pada berbagai bahasa dan mampu memetakan kesamaan makna lintas bahasa ke dalam ruang vektor yang konsisten., seperti kode perintah pada Gambar 10.

```
embedding_model = SentenceTransformer("paraphrase-multilingual-MiniLM-L12-v2")
```

Gambar 10. Pemodelan *Embedding*

5.2. *Generate Embedding*

Tahap *Generate Embedding* merupakan proses transformasi dokumen teks ke dalam bentuk vektor numerik menggunakan model *SentenceTransformer* yang telah ditentukan. Pada tahap ini, setiap dokumen berita diproses oleh model *embedding* untuk menghasilkan representasi vektor berdimensi tetap yang mencerminkan makna semantik teks secara kontekstual. *Embedding* yang dihasilkan memungkinkan dokumen dengan makna serupa memiliki jarak vektor yang lebih dekat dalam ruang semantik, sehingga hubungan antar dokumen dapat dianalisis secara kuantitatif, seperti pada Gambar 11.

```
embeddings = embedding_model.encode(
    texts,
    show_progress_bar=True
)

✓ 57.4s
Batches: 100% 35/35 [00:57<00:00, 1.42s/it]
```

Gambar 11. Generate *Embedding*

5.3. Reduksi Dimensi UMAP

Tahap Reduksi Dimensi dengan *Uniform Manifold Approximation and Projection* (UMAP) dilakukan untuk menurunkan dimensi vektor *embedding* yang semula berdimensi tinggi ke dalam ruang berdimensi lebih rendah dengan tetap mempertahankan struktur semantik data. UMAP memodelkan hubungan kedekatan antar dokumen berdasarkan manifold data, sehingga dokumen yang memiliki kemiripan makna akan tetap berada pada jarak yang relatif dekat setelah proses reduksi dimensi, seperti pada Gambar 12.

```
import umap.umap_ as umap

umap_model = umap.UMAP(
    n_neighbors=15,
    n_components=10,
    min_dist=0.0,
    metric="cosine",
    random_state=42
)
```

Gambar 12. Reduksi Dimensi UMAP

5.4. Clustering dengan HDBSCAN

Tahap *Clustering* dengan *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) bertujuan untuk mengelompokkan dokumen berdasarkan kedekatan vektor hasil reduksi dimensi UMAP. HDBSCAN merupakan algoritma *density-based clustering* yang mampu mengidentifikasi klaster dengan bentuk dan ukuran yang bervariasi tanpa harus menentukan jumlah klaster di awal. HDBSCAN membentuk struktur klaster secara hierarkis berdasarkan kepadatan data, sehingga klaster yang dihasilkan memiliki tingkat koherensi semantik yang lebih baik. Dokumen yang memiliki kemiripan makna tinggi akan cenderung tergabung dalam klaster yang sama, sementara dokumen dengan konteks yang lemah atau ambigu dapat terklasifikasi sebagai outlier. Hasil *clustering* ini menjadi dasar penting dalam pembentukan topik pada tahap

BERTopic, karena setiap klaster merepresentasikan kumpulan dokumen yang memiliki tema yang relatif serupa. Visualisasi sebaran klaster hasil HDBSCAN ditampilkan pada Gambar 13.

```
from hdbscan import HDBSCAN

hdbscan_model = HDBSCAN(
    min_cluster_size=10,
    min_samples=5,
    metric="euclidean",
    cluster_selection_method="leaf",
    prediction_data=True
)
```

Gambar 13. Reduksi Dimensi UMAP

6. MODEL BERTOPIC

Tahap Model BERTopic merupakan inti dari proses pemodelan topik dalam penelitian ini, yang mengintegrasikan *embedding* semantik, reduksi dimensi, dan clustering untuk mengekstraksi topik secara otomatis dari korpus teks. BERTopic bekerja dengan memanfaatkan hasil clustering dokumen untuk membentuk kelompok topik, kemudian mengekstraksi kata-kata representatif pada setiap klaster menggunakan pendekatan *class-based* TF-IDF (c-TF-IDF). Pendekatan ini memungkinkan identifikasi kata kunci yang paling mencerminkan karakteristik unik dari setiap topik, sehingga topik yang dihasilkan bersifat lebih koheren dan mudah diinterpretasikan. Struktur awal pembentukan topik menggunakan model BERTopic, seperti pada Gambar 14.

```
from bertopic import BERTopic

topic_model = BERTopic(
    embedding_model=embedding_model,
    umap_model=umap_model,
    hdbscan_model=hdbscan_model,
    verbose=True
)
```

Gambar 14. Model BERTopic

6.1. Fit-Transform Model

Fit-Transform Model merupakan proses pelatihan dan penerapan model BERTopic terhadap korpus teks penelitian. Pada tahap ini, model melakukan proses fit untuk mempelajari struktur data berdasarkan *embedding* dokumen, hasil reduksi dimensi, dan

pengelompokan klaster, kemudian dilanjutkan dengan proses transform untuk memetakan setiap dokumen ke dalam topik tertentu. Hasil dari tahap ini berupa label topik untuk setiap dokumen serta nilai probabilitas yang menunjukkan tingkat keterkaitan dokumen terhadap topik yang terbentuk, seperti pada Gambar 15.

```
topics, probs = topic_model.fit_transform(texts)
df["Topic"] = topics

✓ 1m 15.2s

Batches: 100% 35/35 [01:02<00:00, 1.45s/it]

2025-12-13 09:11:17,993 - BERTopic - Transformed documents to Embeddings
2025-12-13 09:11:29,052 - BERTopic - Reduced dimensionality
2025-12-13 09:11:29,118 - BERTopic - Clustered reduced embeddings
```

Gambar 15. *Fit-Transform Model*

6.2. Hasil Topik

Hasil topik yang diperoleh dari model BERTopic merepresentasikan kumpulan dokumen yang memiliki kesamaan makna dan konteks secara semantik. Setiap topik dibentuk berdasarkan klaster dokumen hasil proses *fit-transform*, sehingga topik yang dihasilkan mencerminkan pola tematik yang dominan dalam korpus penelitian. Model tidak hanya mengidentifikasi topik utama, tetapi juga mengukur jumlah dokumen yang tergolong dalam setiap topik, sehingga memberikan gambaran mengenai tingkat dominasi dan persebaran isu yang dianalisis, seperti pada Gambar 16.

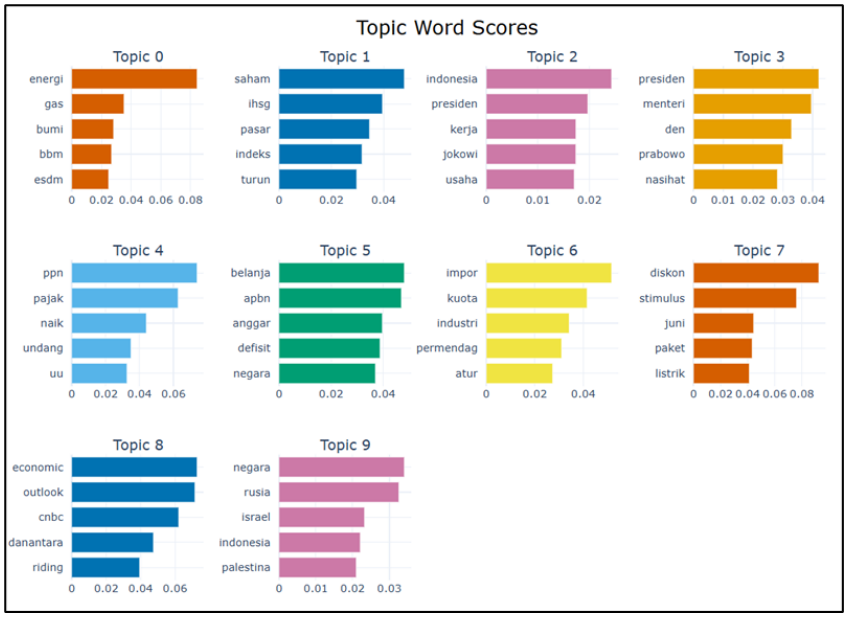
```
topic_info = topic_model.get_topic_info()
print(topic_info.head(10))

✓ 0.0s
```

	Topic	Count	Name
0	-1	430	-1_ekonomi_indonesia_perintah_negara
1	0	65	0_energi_gas_bumi_bbm
2	1	65	1_saham_ihsg_pasar_indeks
3	2	50	2_indonesia_presiden_kerja_jokowi
4	3	50	3_presiden_menteri_den_prabowo
5	4	42	4_ppn_pajak_naik_undang
6	5	35	5_belanja_apbn_anggar_defisit
7	6	33	6_impor_kuota_industri_permendag
8	7	28	7_diskon_stimulus_juni_paket
9	8	28	8_economic_outlook_cnbc_danantara

Gambar 16. Hasil BERTopic

6.3. Visulisasi Topik

Perintah *fig = topic_model.visualize_topics()* digunakan untuk menghasilkan visualisasi interaktif topik yang dibentuk oleh model BERTopic. Visualisasi ini menampilkan sebaran topik dalam ruang dua dimensi berdasarkan hasil reduksi dimensi *embedding*, sehingga hubungan kedekatan antar topik dapat diamati secara visual. Setiap titik merepresentasikan satu topik, sementara jarak antar titik mencerminkan tingkat kemiripan semantik antar topik tersebut. Visualisasi ini membantu dalam memahami struktur global topik serta mengidentifikasi topik yang saling berdekatan atau tumpang tindih seperti pada Gambar 17.



Gambar 17. *Visualisasi Topic Word Scores*

7. PERBANDINGAN LDA DAN BERTOPIC

Berdasarkan hasil evaluasi, dapat disimpulkan bahwa BERTopic menunjukkan performa yang lebih unggul dibandingkan LDA dalam mengekstraksi topik kebijakan ekonomi Presiden Prabowo Subianto. Hal ini ditunjukkan oleh nilai *Coherence Score (c_v)* sebesar 0,72, yang lebih tinggi dibandingkan LDA dengan nilai 0,41, serta kemampuan BERTopic dalam menghasilkan klaster topik yang lebih stabil dan kontekstual.

Sementara itu, model LDA telah mencapai tingkat generalisasi yang cukup baik, namun masih mengalami keterbatasan dalam memisahkan topik yang memiliki kosakata

saling tumpang tindih. Oleh karena itu, BERTopic dinilai lebih sesuai untuk analisis kebijakan publik berbasis teks berita, sedangkan LDA berperan sebagai baseline statistik pembanding. Tabel 2 Ini menunjukkan hasil evaluasi dari kedua metode sebagai berikut.

Tabel 2. Parameter Evaluasi

No.	Parameter Evaluasi	LDA	BERTopic	Keterangan Akademik
1	Jumlah Dokumen	1.000 berita	1.000 berita	Dataset berita ekonomi
2	Metode Representasi	*TF-IDF + BoW*	*Sentence Embedding (IndoBERT)*	-
3	Jumlah Topik Optimal	8 topik	9 topik (+ outlier -1)	Berdasarkan hasil pemodelan
4	Coherence Score (c_v)	0,41	0,72	Semakin tinggi semakin koheren
5	Nilai TF-IDF Dominan	0,031 – 0,084	–	Digunakan pada LDA
6	*Topic Word Scores*	–	0,42 – 0,78	Skor relevansi kata BERTopic
7	Kualitas Pemisahan Topik	Sedang	Tinggi	Dilihat dari overlap kata
8	Overlap Antar Topik	Cukup tinggi	Rendah	BERTopic lebih diskriminatif
9	Kemampuan Konteks Semantik	Rendah–Sedang	Sangat Tinggi	BERT menangkap konteks
10	*Clustering Hierarkis*	Tidak ada	Ada (HDBSCAN)	-
11	Stabilitas Klaster	Sensitif jumlah topik	Stabil otomatis	HDBSCAN adaptif
12	Interpretabilitas Topik	Cukup	Sangat Baik	Evaluasi kualitatif
13	Visualisasi	*PyLDAvis*	*Intertopic Distance Map*	-
14	Relevansi Isu Kebijakan	Cukup r*epresentatif*	Sangat representatif	Topik ekonomi lebih spesifik
15	Waktu Komputasi	Lebih cepat	Lebih lambat	*Trade-off performa*
16	Model Lebih Unggul	–	BERTopic	Berdasarkan evaluasi keseluruhan

8. KESIMPULAN

Pemodelan topik menggunakan LDA dan BERTopic berhasil mengidentifikasi isu utama kebijakan ekonomi Presiden Prabowo Subianto berdasarkan data berita ekonomi. Topik-topik yang dihasilkan mencerminkan fokus kebijakan publik yang dominan, seperti kebijakan fiskal, sektor energi, pasar modal, dan stimulus ekonomi, sehingga

menunjukkan bahwa pendekatan *topic modeling* efektif digunakan untuk mengekstraksi wacana kebijakan dari data teks berskala besar.

Model BERTopic menunjukkan kinerja yang lebih unggul dibandingkan LDA dalam hal koherensi dan interpretabilitas topik. Hal ini ditunjukkan oleh nilai *Coherence Score* (c_v) yang lebih tinggi pada BERTopic (0,72) dibandingkan LDA (0,61), serta kemampuan BERTopic dalam menghasilkan label topik yang lebih kontekstual dan semantik melalui pemanfaatan *embedding* IndoBERT dan *hierarchical clustering* berbasis HDBSCAN.

Model LDA tetap memiliki nilai metodologis sebagai pendekatan *baseline* berbasis *probabilistik* yang dapat digunakan untuk memberikan gambaran awal struktur topik dalam korpus. Namun, keterbatasan LDA dalam menangkap konteks semantik menyebabkan adanya tumpang tindih kosakata antar topik, sehingga kurang optimal untuk analisis kebijakan publik yang membutuhkan pemahaman makna secara lebih mendalam.

DAFTAR PUSTAKA

[1] A. C. L and O. A. Putri, “Analisis Badai Inflasi Hipotetis : Dampak dan Respons Kebijakan pada Awal Pemerintahan Prabowo Subianto di Indonesia Tahun 2025,” pp. 187–202, 2025.

[2] S. I. Ishak, O. Arnilia, T. Widodo, I. G. Nyoman, and A. Bisma, “Analisis sentimen terhadap pemerintahan Prabowo – Gibran menggunakan IndoBERT dan LDA,” vol. 7, no. 2, pp. 72–82, 2025, doi: 10.37905/jji.v1i2.34895.

[3] Fadisa Rahma Devi Triana and Muhammad Thoyib Amali, “Analisis Framing Pemberitaan Program Kerja Makan Siang Gratis Prabowo-Gibran Dalam Media Online Liputan6.Com Dan Republika.co.id,” *J. Ilmu Sos. dan Ilmu Polit.*, vol. 13, no. 3, pp. 603–614, 2024, [Online]. Available: www.publikasi.unitri.ac.id

[4] Z. Tang, X. Pan, and Z. Gu, “Analyzing public demands on China’s online government inquiry platform: A BERTopic-Based topic modeling study,” *PLoS One*, vol. 19, no. 2 February, pp. 1–26, 2024, doi: 10.1371/journal.pone.0296855.

[5] T. Gokcimen and B. Das, "Exploring climate change discourse on social media and blogs using a topic modeling analysis," *Heliyon*, vol. 10, no. 11, p. e32464, 2024, doi: 10.1016/j.heliyon.2024.e32464.

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[7] H. Tommy *et al.*, "KLASTERISASI BERITA BAHASA INDONESIA DENGAN MENGGUNAKAN K- MEANS DAN WORD EMBEDDING CLUSTERING INDONESIA NEWS USING K-MEANS AND WORD EMBEDDING," vol. 10, no. 3, 2023, doi: 10.25126/jtiik.2023106468.

[8] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *COLING 2020 - 28th Int. Conf. Comput. Linguist. Proc. Conf.*, pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.