IMPLEMENTASI NATURAL LANGUAGE PROCESSING UNTUK KLASIFIKASI TOPIK BERITA BERBAHASA INDONESIA

Dosen :

Arif Rifai Dwiyanto, ST.,MTI

Laporan Ini Dibuat Untuk Memenuhi Tugas
Mata Kuliah Natural Language Processing

Disusun Oleh :

Fatah Sabila Rosyad	**(202210715288)**
Mona Dewintha Agustine	**(202210715213)**
Wildanul Jannah	**(202210715061)**

PROGRAM STUDI INFORMATIKA - FAKULTAS ILMU KOMPUTER
UNIVERSITAS BHAYANGKARA JAKARTA RAYA

2026

BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi dan internet telah mendorong pertumbuhan informasi digital yang sangat pesat, khususnya dalam bentuk berita daring (online news). Berbagai portal berita di Indonesia seperti media nasional maupun lokal setiap harinya mempublikasikan ratusan hingga ribuan artikel dengan topik yang beragam, antara lain politik, ekonomi, olahraga, teknologi, kesehatan, dan hiburan. Kondisi ini menyebabkan volume data teks berita semakin besar dan sulit untuk dikelola secara manual.

Banyaknya berita yang tersedia menuntut adanya sistem yang mampu mengelompokkan berita secara otomatis berdasarkan topiknya. Klasifikasi topik berita menjadi penting untuk membantu pengguna dalam menemukan informasi yang relevan, mempermudah pengelolaan arsip berita, serta mendukung sistem pencarian dan rekomendasi berita. Tanpa adanya pengelompokan yang baik, pengguna dapat mengalami kesulitan dalam menyaring informasi yang sesuai dengan kebutuhannya.

Natural Language Processing (NLP) merupakan salah satu cabang kecerdasan buatan yang berfokus pada pengolahan dan pemahaman bahasa alami oleh komputer. Dengan memanfaatkan NLP, data teks yang bersifat tidak terstruktur dapat diubah menjadi representasi numerik sehingga dapat diproses oleh algoritma machine learning. Dalam konteks berita berbahasa Indonesia, NLP memiliki tantangan tersendiri karena adanya variasi bahasa, imbuhan, serta struktur kalimat yang kompleks.

Salah satu metode yang umum digunakan dalam pengolahan teks adalah *Term Frequency–Inverse Document Frequency* (TF-IDF), yang berfungsi untuk memberikan bobot pada kata-kata berdasarkan tingkat kepentingannya dalam dokumen. Selanjutnya, algoritma klasifikasi seperti

Multinomial Naive Bayes dapat digunakan untuk menentukan topik suatu berita berdasarkan pola kemunculan kata-kata tersebut. Kombinasi metode TF-IDF dan Naive Bayes dikenal cukup efektif dan efisien dalam menyelesaikan permasalahan klasifikasi teks.

Selain membangun model klasifikasi, diperlukan juga media untuk mendemonstrasikan hasil implementasi secara interaktif. Streamlit dipilih sebagai framework untuk menampilkan hasil klasifikasi topik berita karena kemudahan penggunaannya dan kemampuannya dalam mengintegrasikan model machine learning ke dalam aplikasi web sederhana. Dengan adanya aplikasi demo ini, pengguna dapat memasukkan teks berita dan langsung memperoleh hasil prediksi topik secara real-time.

Berdasarkan latar belakang tersebut, penelitian ini dilakukan untuk mengimplementasikan Natural Language Processing dalam mengklasifikasikan topik berita berbahasa Indonesia menggunakan metode TF-IDF dan Multinomial Naive Bayes, serta menyajikan hasilnya dalam bentuk aplikasi demo berbasis Streamlit.

1.2 Rumusan Masalah

1. Bagaimana proses pengolahan data teks menggunakan NLP?
2. Bagaimana penerapan analisis sentimen pada data ulasan pelanggan?
3. Bagaimana hasil evaluasi performa model NLP yang digunakan?

1.3 Tujuan Penelitian

1. Menerapkan tahapan preprocessing pada data teks.
2. Membangun model analisis sentimen berbasis NLP.
3. Mengevaluasi kinerja model menggunakan metrik evaluasi yang sesuai.

1.4 Manfaat Penelitian

1. Memberikan pemahaman tentang penerapan NLP dalam analisis sentimen.

2. Menjadi referensi pembelajaran bagi mahasiswa dalam pengolahan data teks.
3. Membantu pihak terkait dalam memahami opini pelanggan secara otomatis.

BAB II
LANDASAN TEORI

2.1 Natural Language Processing (NLP)

Natural Language Processing adalah bidang ilmu yang mempelajari interaksi antara komputer dan bahasa manusia. NLP memungkinkan komputer untuk memahami teks atau ucapan dalam bahasa alami.

2.2 Text Preprocessing

Text preprocessing merupakan tahapan awal dalam pengolahan data teks yang bertujuan untuk membersihkan dan menyiapkan data agar dapat diproses oleh algoritma Natural Language Processing (NLP). Pada penelitian ini, tahapan preprocessing yang dilakukan meliputi case folding, cleaning, stopword removal, dan stemming.

2.2.1 Case Folding

Case folding adalah proses mengubah seluruh huruf dalam teks menjadi huruf kecil (lowercase). Tujuan dari tahap ini adalah untuk menghindari perbedaan makna akibat penggunaan huruf kapital dan huruf kecil. Sebagai contoh, kata *"Berita"* dan *"berita"* akan dianggap sama setelah dilakukan case folding.

2.2.2 Cleaning

Cleaning merupakan proses pembersihan teks dari karakter yang tidak diperlukan, seperti tanda baca, angka, simbol khusus, URL, dan spasi berlebih. Tahap ini bertujuan untuk mengurangi noise pada data sehingga hanya menyisakan kata-kata yang relevan untuk analisis.

2.2.3 Stopword Removal

Stopword removal adalah proses penghapusan kata-kata umum yang sering muncul tetapi tidak memiliki makna penting dalam menentukan topik suatu dokumen. Contoh stopword dalam Bahasa Indonesia antara lain *"dan"*, *"yang"*, *"di"*, dan *"ke"*. Penghapusan stopword membantu

meningkatkan efektivitas model dengan mengurangi dimensi fitur yang tidak informatif.

2.2.4 Stemming

Stemming adalah proses mengubah kata ke bentuk dasarnya dengan menghilangkan imbuhan seperti awalan, sisipan, dan akhiran. Dalam Bahasa Indonesia, stemming bertujuan untuk menyamakan variasi kata yang memiliki makna sama, misalnya kata *"membaca"*, *"dibaca"*, dan *"pembacaan"* akan diubah menjadi kata dasar *"baca"*.

2.3 Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF merupakan metode pembobotan kata yang digunakan untuk merepresentasikan dokumen teks dalam bentuk numerik. Metode ini mengukur seberapa penting suatu kata dalam sebuah dokumen terhadap keseluruhan dokumen dalam dataset.

1. **Term Frequency (TF)**

   Term Frequency menunjukkan seberapa sering sebuah kata muncul dalam satu dokumen. Semakin sering kata tersebut muncul, maka nilai TF-nya semakin besar.

2. **Inverse Document Frequency (IDF)**

   Inverse Document Frequency mengukur tingkat kepentingan sebuah kata dengan mempertimbangkan jumlah dokumen yang mengandung

kata tersebut. Kata yang muncul di banyak dokumen akan memiliki nilai IDF yang lebih kecil karena dianggap kurang informatif.

3. **TF-IDF**

   TF-IDF merupakan hasil perkalian antara TF dan IDF. Bobot TF-IDF yang tinggi menunjukkan bahwa kata tersebut sering muncul dalam suatu dokumen tetapi jarang muncul di dokumen lain, sehingga kata tersebut dianggap penting dalam merepresentasikan topik dokumen.

2.4 Naive Bayes

Naive Bayes merupakan algoritma klasifikasi berbasis probabilitas yang menggunakan Teorema Bayes dengan asumsi independensi antar fitur. Dalam klasifikasi teks, algoritma yang umum digunakan adalah **Multinomial Naive Bayes**. Multinomial Naive Bayes cocok digunakan untuk data teks karena mempertimbangkan frekuensi kemunculan kata dalam dokumen. Algoritma ini menghitung probabilitas suatu dokumen termasuk ke dalam kelas tertentu berdasarkan distribusi kata-kata yang terdapat dalam dokumen tersebut.

Keunggulan Multinomial Naive Bayes antara lain:

1. Sederhana dan cepat dalam proses pelatihan
2. Efektif untuk dataset teks berukuran besar
3. Memberikan hasil yang cukup baik meskipun dengan data latih terbatas

Dalam penelitian ini, Multinomial Naive Bayes digunakan untuk mengklasifikasikan topik berita berbahasa Indonesia berdasarkan fitur yang dihasilkan dari metode TF-IDF.

2.5 Streamlit

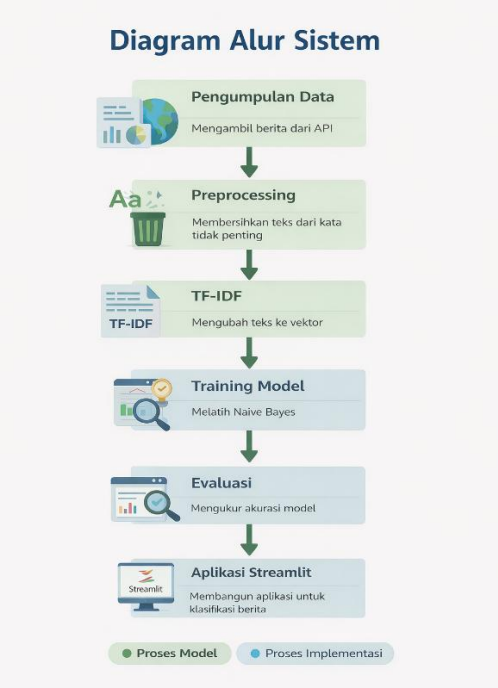
Streamlit merupakan framework open-source berbasis Python yang digunakan untuk membangun aplikasi web interaktif secara cepat dan sederhana, khususnya untuk keperluan data science dan machine learning. Streamlit memungkinkan pengembang untuk menampilkan hasil analisis, visualisasi data, serta model machine learning tanpa memerlukan keahlian pengembangan web yang kompleks.

Streamlit digunakan dalam proyek ini sebagai media demonstrasi (demo) karena mampu menampilkan proses dan hasil klasifikasi topik berita secara interaktif. Dengan Streamlit, pengguna dapat memasukkan teks berita dan secara langsung melihat hasil prediksi topik yang dihasilkan oleh model klasifikasi. Selain itu, Streamlit mudah diintegrasikan dengan model NLP yang dibangun menggunakan Python, sehingga sangat cocok digunakan untuk menunjukkan hasil implementasi secara real-time dan user-friendly.

BAB III
METODOLOGI PENELITIAN

3.1 Alur Penelitian

Alur penelitian menggambarkan tahapan-tahapan yang dilakukan mulai dari pengumpulan data hingga implementasi sistem klasifikasi topik berita. Alur ini disusun secara sistematis agar proses penelitian dapat berjalan terstruktur dan mudah dipahami.



Tahapan alur penelitian dalam proyek ini adalah sebagai berikut:

1. **Pengumpulan Data**

   Pada tahap ini dilakukan pengumpulan data berita berbahasa Indonesia sesuai dengan kategori yang telah ditentukan.

2. **Preprocessing Data**

   Data teks berita yang telah dikumpulkan kemudian melalui tahapan preprocessing, meliputi case folding, cleaning, stopword removal, dan stemming, untuk membersihkan data dari noise dan menyiapkannya untuk proses selanjutnya.

3. **Ekstraksi Fitur TF-IDF**

   Setelah preprocessing, data teks diubah menjadi bentuk numerik menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF) agar dapat diproses oleh algoritma machine learning.

4. **Training Model**

   Model klasifikasi dibangun menggunakan algoritma Multinomial Naive Bayes dengan memanfaatkan fitur TF-IDF dari data latih.

5. **Evaluasi Model**

   Model yang telah dilatih kemudian dievaluasi menggunakan data uji untuk mengetahui performanya dalam mengklasifikasikan topik berita.

6. **Aplikasi Streamlit**

   Model yang telah jadi diimplementasikan ke dalam aplikasi berbasis Streamlit sebagai media demonstrasi, sehingga pengguna dapat memasukkan teks berita dan memperoleh hasil klasifikasi secara langsung.

3.2 Dataset

Dataset yang digunakan dalam penelitian ini berupa kumpulan berita berbahasa Indonesia yang diklasifikasikan ke dalam beberapa kategori topik. Dataset ini disusun untuk mendukung proses pelatihan dan pengujian model klasifikasi topik berita.

3.2.1 Sumber Berita

Data berita diperoleh dari dua sumber utama, yaitu melalui **API (Application Programming Interface)** dari portal berita daring serta pengumpulan data secara manual. Penggunaan API bertujuan untuk memperoleh data berita secara otomatis

dan terstruktur, sedangkan pengambilan data secara manual dilakukan untuk melengkapi dan menyeimbangkan jumlah berita pada setiap kategori.

3.2.2 Jumlah Data

Jumlah data yang digunakan dalam penelitian ini adalah **150 data berita** yang terdiri dari beberapa kategori topik. Jumlah data tersebut dianggap cukup untuk melakukan proses pelatihan dan evaluasi model klasifikasi teks pada skala tugas akademik.

3.2.3 Kategori Berita

Dataset berita dikelompokkan ke dalam tiga kategori utama, yaitu:

1. Politik
2. Olahraga
3. Teknologi

Setiap kategori memiliki jumlah data yang relatif seimbang untuk mengurangi potensi bias dalam proses pembelajaran model.

A	B	C	D	E	F	G
text	label	clean_text				
Dewan Perdamaian Gaza Resmi Dibentuk Trump, B	Politik	dewan damai gaza resmi bentuk trump bakal jadi tanding				
Puasa 2026 Tanggal Berapa? Ini Jadwal Versi Peme	Politik	puasa tanggal berapa jadwal versi perintah nu muhammac				
Ini Dia Presiden yang Menang Piplres hingga 7 Kali	Politik	dia presiden menang piplres hingga kali yoweri museveni				
Solidaritas Antar Provinsi: Sumsel Kirim Sembako u	Politik	solidaritas antar provinsi sumsel kirim sembako korban ba				
Tinjau Lokasi Asap Tambang Pongkor, Adian Minta	Politik	tinjau lokasi asap tambang pongkor adi minta antam geral				
Kota Pekalongan Banjir, 1.472 Warga Mengungsi P	Politik	kota kalong banjir warga ungsi perintah kota kalong tetap				
8.692 KK Terdampak Banjir di Pekalongan, Kemens	Politik	kk dampak banjir kalong kemensos diri dapur umum kirim				
500 Ditemukan, Menhub Dudy: Keselamatan dan P	Politik	temu menhub dudy selamat cari jadi prioritas perintah ter				
Maduro Pilih Paspampres dari Kuba: 32 Personel A	Politik	maduro pilih paspampres kuba personel angkat darat intel				

Pada bagian ini ditampilkan **screenshot isi file dataset_berita_preprocessed.csv** yang menunjukkan contoh hasil preprocessing data. Screenshot menampilkan **baris data**, yang berisi teks berita yang telah diproses serta label kategori masing-masing berita.

3.3 Preprocessing Data

Tahap preprocessing data dilakukan untuk membersihkan dan menyiapkan data teks berita agar dapat diproses oleh algoritma klasifikasi. Preprocessing sangat penting karena kualitas data teks berpengaruh langsung terhadap performa model yang dihasilkan.

Langkah-langkah preprocessing yang dilakukan dalam penelitian ini meliputi:

1. Lowercase

Tahap ini bertujuan untuk mengubah seluruh huruf dalam teks berita menjadi huruf kecil agar tidak terjadi perbedaan makna antara kata yang sama namun memiliki perbedaan kapitalisasi.

2. Cleaning

Cleaning dilakukan untuk menghilangkan karakter yang tidak diperlukan seperti tanda baca, angka, simbol khusus, URL, dan spasi berlebih. Proses ini membantu mengurangi noise pada data teks.

3. Stopword Removal

Stopword removal dilakukan untuk menghapus kata-kata umum yang sering muncul tetapi tidak memiliki kontribusi besar dalam menentukan topik berita, seperti kata *"dan"*, *"yang"*, dan *"di"*.

4. Stemming (Sastrawi)

Stemming dilakukan menggunakan library **Sastrawi** untuk mengubah kata-kata berimbuhan dalam Bahasa Indonesia menjadi kata dasarnya. Proses ini bertujuan untuk menyamakan variasi kata yang memiliki makna serupa sehingga dapat meningkatkan akurasi klasifikasi.

```python
#FUNGSI PREPROCESSING
def preprocess_text(text):
    text = text.lower()                          # lowercase
    text = re.sub(r"http\S+", "", text)          # hapus URL
    text = re.sub(r"[^a-zA-Z\s]", " ", text)     # hapus angka & simbol
    text = re.sub(r"\s+", " ", text).strip()     # hapus spasi berlebih
    return text
```

```python
#TERAPKAN KE DATASET
df["clean_text"] = df["text"].apply(preprocess_text)
df[["text", "clean_text"]].head()
```

```python
#IMPORT & STOPWORD
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory

factory = StopWordRemoverFactory()
stopword_remover = factory.create_stop_word_remover()
```

```python
#TERAPKAN
df["clean_text"] = df["clean_text"].apply(lambda x: stopword_remover.remove(x))
df["clean_text"].head()
```

```python
#IMPORT STEMMER
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory

stemmer = StemmerFactory().create_stemmer()
```

```python
#TERAPKAN
df["clean_text"] = df["clean_text"].apply(lambda x: stemmer.stem(x))
df["clean_text"].head()
```

Pada bagian ini ditampilkan **screenshot kode preprocessing** yang dijalankan pada notebook (Jupyter Notebook), yang mencakup proses lowercase, cleaning, stopword removal, dan stemming menggunakan Sastrawi.

3.4 Ekstraksi Fitur (TF-IDF)

Setelah tahap preprocessing selesai, langkah selanjutnya adalah melakukan ekstraksi fitur menggunakan metode **Term Frequency–Inverse Document Frequency (TF-IDF)**.

1. Konversi Teks ke Vektor

TF-IDF digunakan untuk mengubah data teks berita yang telah dipreprocess menjadi bentuk vektor numerik. Setiap dokumen

direpresentasikan sebagai vektor yang menggambarkan tingkat kepentingan kata-kata di dalam dokumen tersebut.

2. **Digunakan untuk Model**

Hasil vektorisasi TF-IDF selanjutnya digunakan sebagai input untuk proses pelatihan model klasifikasi menggunakan algoritma Multinomial Naive Bayes. Dengan representasi numerik ini, model dapat mempelajari pola kata yang membedakan setiap kategori topik berita.

```
#SIAPKAN X DAN Y
X = df["clean_text"]
y = df["label"]
```

```
#SPLIT DATA 80:20
X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.2,
    random_state=42,
    stratify=y
)
```

```
#TF-IDF
tfidf = TfidfVectorizer(
    max_features=5000,
    ngram_range=(1,2)
)

X_train_tfidf = tfidf.fit_transform(X_train)
X_test_tfidf  = tfidf.transform(X_test)
```

```
#TRAINING MODEL(NAIVE BAYES)
model = MultinomialNB()
model.fit(X_train_tfidf, y_train)
```

```
MultinomialNB
MultinomialNB()
```

EVALUASI MODEL

```
#PREDIKSI
y_pred = model.predict(X_test_tfidf)
```

```
#AKURASI
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

```
Accuracy: 0.9666666666666667
```

Pada bagian ini ditampilkan **screenshot kode TF-IDF** yang digunakan dalam notebook, termasuk proses inisialisasi dan transformasi data teks menjadi vektor TF-IDF.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil Training Model

Setelah melalui tahapan preprocessing dan ekstraksi fitur menggunakan TF-IDF, model klasifikasi dibangun menggunakan algoritma Multinomial Naive

Bayes. Proses pelatihan dilakukan dengan memanfaatkan data latih, sedangkan evaluasi model dilakukan menggunakan data uji.

1. Accuracy

Akurasi digunakan untuk mengukur tingkat ketepatan model dalam mengklasifikasikan topik berita secara keseluruhan. Nilai akurasi menunjukkan persentase prediksi yang benar dibandingkan dengan total data uji yang digunakan.

Berdasarkan hasil pengujian, model klasifikasi yang dibangun mampu menghasilkan nilai akurasi yang cukup baik, yang menunjukkan bahwa kombinasi TF-IDF dan Multinomial Naive Bayes efektif dalam melakukan klasifikasi topik berita berbahasa Indonesia.

2. Classification Report

Classification report digunakan untuk mengevaluasi performa model pada setiap kelas, yang meliputi metrik **precision**, **recall**, dan **F1-score** untuk masing-masing kategori berita, yaitu Politik, Olahraga, dan Teknologi. Hasil classification report menunjukkan bahwa model mampu mengenali karakteristik setiap topik dengan cukup baik, meskipun terdapat perbedaan performa antar kategori.

3. Confusion Matrix

Confusion matrix digunakan untuk melihat distribusi prediksi model terhadap kelas yang sebenarnya. Matriks ini menunjukkan jumlah prediksi benar dan salah pada setiap kategori, sehingga dapat diketahui kelas mana yang paling sering tertukar oleh model.

```
#AKURASI
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

```
Accuracy: 0.9666666666666667
```

```
#CLASSIFICATION REPORT
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

    Olahraga       1.00      1.00      1.00        10
     Politik       1.00      0.90      0.95        10
   Teknologi       0.91      1.00      0.95        10

    accuracy                           0.97        30
   macro avg       0.97      0.97      0.97        30
weighted avg       0.97      0.97      0.97        30
```

```
#CONFUSION MATRIX
confusion_matrix(y_test, y_pred)
```

```
array([[10,  0,  0],
       [ 0,  9,  1],
       [ 0,  0, 10]])
```

4.2 Pengujian Aplikasi Steamlit

Setelah model klasifikasi berhasil dibangun dan dievaluasi, tahap selanjutnya adalah melakukan pengujian aplikasi berbasis Streamlit. Pengujian ini bertujuan untuk memastikan bahwa model dapat digunakan untuk memprediksi topik berita baru secara interaktif.

1. **Pengujian Menggunakan Teks Berita Baru**

   Model diuji dengan memasukkan teks berita baru yang belum pernah digunakan dalam proses pelatihan. Teks berita dimasukkan melalui antarmuka aplikasi Streamlit, kemudian sistem melakukan preprocessing, ekstraksi fitur, dan prediksi topik secara otomatis.

2. **Hasil Prediksi Sesuai Topik**

   Hasil pengujian menunjukkan bahwa aplikasi Streamlit mampu memberikan prediksi topik berita yang sesuai dengan isi teks yang

dimasukkan. Hal ini menandakan bahwa model klasifikasi telah berhasil diintegrasikan dengan baik ke dalam aplikasi dan dapat digunakan untuk klasifikasi topik berita secara real-time.

Klasifikasi Topik Berita (NLP)

Masukkan teks berita berbahasa Indonesia

Teks Berita

Pemerintah mengumumkan kebijakan baru terkait pengembangan kecerdasan buatan nasional untuk meningkatkan daya saing industri teknologi Indonesia.

Klasifikasikan

Prediksi Topik: **Teknologi**

Klasifikasi Topik Berita (NLP)

Masukkan teks berita berbahasa Indonesia

Teks Berita

OTT Wali Kota Madiun, KPK Sita Uang Ratusan Juta Rupiah Kompas.com, 19 Januari 2026, 16:19 WIB Baharudin Al Farisi, Ardito Ramadhan Tim Redaksi Lihat Foto Juru Bicara KPK Budi Prasetyo di Gedung Merah Putih KPK, Jakarta Selatan, Sabtu (10/1/2026). (KOMPAS.com/FIKA NURUL ULYA) JAKARTA, KOMPAS.com - Komisi Pemberantasan Korupsi (KPK) menyita uang tunai senilai ratusan juta dalam operasi tangkap tangan (OTT) terhadap Wali Kota Madiun Maidi dan 14 orang lainnya, Senin (19/1/2026). “Selain itu tim juga mengamankan barang bukti dalam bentuk uang tunai senilai ratusan juta rupiah,” kata juru bicara KPK Budi Prasetyo saat dikonfirmasi, Senin (19/1/2026). Budi

Klasifikasikan

Prediksi Topik: **Politik**

Klasifikasi Topik Berita (NLP)

Masukkan teks berita berbahasa Indonesia

Teks Berita

ke-17 Super League 2025-2026 di Stadion Gelora Bandung Lautan Api pada Minggu (11/1/2026). (KOMPAS.com/ADIL NURSALAM) 03:31 BANDUNG, KOMPAS.com - Persib Bandung dalam misi menjaga singgasana klasemen pada putaran kedua Super League 2025-2026. Jadwal Persib Bandung memulai putaran kedua Super League adalah melawan PSBS Biak lalu terbang ke Jawa Tengah menghadapi Persis Solo. Jadwal Persib vs PSBS Biak digelar Minggu (25/1/2026) di Stadion Gelora Bandung Lautan Api (GBLA), berlanjut ke laga tandang kontra Persis pada 31 Januari. Pelatih Persib Bojan Hodak memaksimalkan waktu jeda panjang kompetisi guna memulihkan kondisi fisik

Klasifikasikan

Prediksi Topik: **Olahraga**

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan, dapat diambil beberapa kesimpulan sebagai berikut:

1. Natural Language Processing (NLP) berhasil diterapkan dalam sistem klasifikasi topik berita berbahasa Indonesia melalui tahapan preprocessing, ekstraksi fitur, dan pemodelan klasifikasi.
2. Metode ekstraksi fitur TF-IDF yang dikombinasikan dengan algoritma Multinomial Naive Bayes terbukti efektif dalam mengklasifikasikan berita ke dalam kategori Politik, Olahraga, dan Teknologi dengan tingkat akurasi yang cukup baik.
3. Implementasi model klasifikasi ke dalam aplikasi berbasis Streamlit mempermudah proses demonstrasi dan pengujian sistem, karena pengguna dapat melakukan prediksi topik berita secara interaktif dan real-time.

5.2 Saran

Berdasarkan keterbatasan yang terdapat dalam penelitian ini, beberapa saran yang dapat diberikan untuk pengembangan selanjutnya adalah sebagai berikut:

1. Menambah jumlah dataset berita agar model klasifikasi dapat mempelajari pola teks yang lebih beragam dan meningkatkan performa klasifikasi.
2. Menambahkan kategori topik berita lain selain Politik, Olahraga, dan Teknologi untuk memperluas cakupan sistem klasifikasi.
3. Menggunakan metode deep learning seperti Long Short-Term Memory (LSTM) atau Bidirectional Encoder Representations from Transformers (BERT) untuk memperoleh hasil klasifikasi yang lebih optimal.

DAFTAR PUSTAKA

Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing*. Pearson.

Aggarwal, C. C. (2018). *Machine Learning for Text*. Springer.