



## Klasifikasi Berita Bahasa Indonesia Dengan Menggunakan Metode K-Nearest Neighbor Dan Naive Bayes

Komariah Kukum Manieh Nuryasin<sup>1</sup>, Taswanda Taryo<sup>2</sup>, dan Sudarno<sup>3</sup>

Program Studi Teknik Informatika S-2, Universitas Pamulang

Jalan Raya Puspitek No. 46 Buaran, Serpong, Tangerang Selatan, Banten, Indonesia

Email: um.komariah@gmail.com<sup>1</sup>, taswandataryo@gmail.com<sup>2</sup>, dan sudarnowiharjo@gmail.com<sup>3</sup>

### ABSTRACT

*In the era of rapid development of information technology, the need for a news classification system is crucial to manage the increasing volume of information. This study aims to develop a news classification system in Indonesian into five main categories: Politics, Economy, Health, Security, and Poverty. The methods used include the K-Nearest Neighbor (KNN) algorithm and Naïve Bayes. The dataset consists of 2,000 news items obtained from Kaggle, with preprocessing stages including cleaning, tokenizing, normalization, and TF-IDF weighting. The evaluation was carried out through three data sharing scenarios: 70%-30%, 80%-20%, and 90%-10%. The results showed that the KNN algorithm achieved the highest accuracy of 89% in the 80%-20% and 90%-10% scenarios, while Naïve Bayes produced the best accuracy of 78.66% in the 70%-30% scenario. KNN proved to be more reliable for data with balanced category distribution, while Naïve Bayes required further adjustment, especially for underrepresented data categories. This research provides significant contributions to the development of an automatic news classification system, which can be implemented to improve user experience in accessing information.*

**Keywords:** News Classification, K-Nearest Neighbor, Naïve Bayes, Text Mining, TF-IDF.

### ABSTRAK

Dalam era perkembangan teknologi informasi yang pesat, kebutuhan akan sistem klasifikasi berita menjadi krusial untuk mengelola volume informasi yang terus bertambah. Penelitian ini bertujuan mengembangkan sistem klasifikasi berita berbahasa Indonesia ke dalam lima kategori utama: Politik, Ekonomi, Kesehatan, Keamanan, dan Kemiskinan. Metode yang digunakan meliputi algoritma K-Nearest Neighbor (KNN) dan Naïve Bayes. Dataset terdiri dari 2.000 berita yang diperoleh dari Kaggle, dengan tahapan preprocessing meliputi cleaning, tokenizing, normalisasi, dan pembobotan TF-IDF. Evaluasi dilakukan melalui tiga skenario pembagian data: 70%-30%, 80%-20%, dan 90%-10%. Hasil penelitian menunjukkan bahwa algoritma KNN mencapai akurasi tertinggi sebesar 89% pada skenario 80%-20% dan 90%-10%, sedangkan Naïve Bayes menghasilkan akurasi terbaik sebesar 78,66% pada skenario 70%-30%. KNN terbukti lebih andal untuk data dengan distribusi kategori seimbang, sementara Naïve Bayes memerlukan penyesuaian lebih lanjut, terutama pada kategori data yang kurang terwakili. Penelitian ini memberikan kontribusi signifikan dalam pengembangan sistem klasifikasi berita otomatis, yang dapat diimplementasikan untuk meningkatkan pengalaman pengguna dalam mengakses informasi.

**Kata Kunci:** Klasifikasi Berita, K-Nearest Neighbor, Naïve Bayes, Text Mining, TF-IDF.

## 1. PENDAHULUAN

Dalam era perkembangan teknologi informasi dan internet yang semakin pesat, kebutuhan akan sistem klasifikasi berita yang efisien menjadi sangat penting. Volume berita yang dihasilkan secara online terus meningkat, sehingga pembaca membutuhkan sistem yang dapat membantu menemukan informasi relevan dengan cepat [1].

Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi berita berbahasa Indonesia menggunakan algoritma K-Nearest Neighbor (KNN) dan Naive Bayes. Klasifikasi dilakukan pada lima kategori berita, yaitu Politik, Ekonomi, Kesehatan, Keamanan, dan Kemiskinan.

Klasifikasi adalah penyusunan bersistem dalam kelompok atau golongan menurut kaidah atau standar yang ditetapkan. Klasifikasi adalah proses penemuan model (atau fungsi) yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui[2].

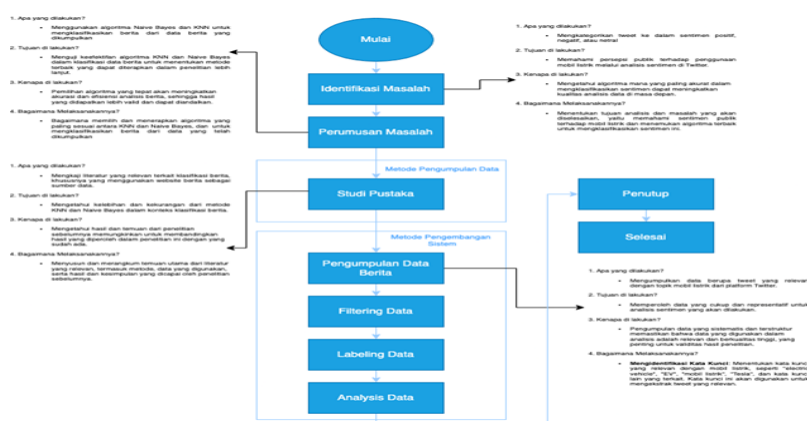
Klasifikasi berita adalah proses mengelompokkan dokumen berita ke dalam kategori yang telah ditentukan, seperti Ekonomi,Politik,Kesehatan,Keamanan dan Kemiskinan. Proses ini tidak hanya membantu dalam pengelolaan informasi, tetapi juga meningkatkan pengalaman pengguna dalam mengakses berita[3]. Salah satu metode yang populer dalam klasifikasi adalah K-Nearest Neighbor (KNN). Metode ini bekerja dengan cara mencari k tetangga terdekat dari data yang ingin diklasifikasikan dan menentukan kategori berdasarkan mayoritas kategori dari tetangga tersebut[4].

Klasifikasi dokumen adalah sebuah metode perolehan informasi yang untuk menentukan atau mengkategorikan suatu dokumen ke dalam satu atau lebih kelompok[5]. K-Nearest Neighbor (KNN) adalah suatu metode yang menggunakan algoritma supervised dimana data testing yang baru diklasifikasikan berdasarkan mayoritas kelas pada KNN[6]. Text mining adalah proses ekstraksi informasi penting dari teks tidak terstruktur dengan mengidentifikasi pola atau hubungan di dalam data. Tahapan text mining meliputi parsing, penghilangan fitur yang tidak relevan, dan analisis data untuk menghasilkan wawasan yang berguna[7].

Naive Bayes adalah algoritma berbasis probabilitas yang menggunakan Teorema Bayes. Algoritma ini mengasumsikan bahwa setiap fitur bersifat independen, sehingga perhitungannya menjadi lebih sederhana dan cepat[8]. Naive Bayes sangat cocok untuk klasifikasi teks. TF-IDF (Term Frequency-Inverse Document Frequency) adalah metode pembobotan kata yang mengukur seberapa penting sebuah kata dalam dokumen tertentu relatif terhadap kumpulan dokumen[9]. Metode ini meningkatkan akurasi klasifikasi dengan menyoroti kata-kata kunci.

## **2. METODE**

Penelitian ini menggunakan dataset berisi 2.000 berita berbahasa Indonesia yang diperoleh dari situs Kaggle[10]. Tahapan penelitian meliputi pengumpulan data dengan mengkategorikan berita ke dalam lima kategori utama; preprocessing yang mencakup cleaning, case folding, tokenizing, normalisasi, stopwords removal, dan stemming; pembobotan kata menggunakan metode TF-IDF (Term Frequency-Inverse Document Frequency) untuk meningkatkan relevansi klasifikasi; klasifikasi menggunakan algoritma KNN, yang bekerja dengan mencari k tetangga terdekat dan menentukan kategori berdasarkan mayoritas kelas tetangga, dan Naive Bayes, algoritma probabilistik yang menghitung peluang kategori berdasarkan distribusi data dalam dokumen; serta evaluasi akurasi menggunakan pembagian data latih dan uji dengan skenario 80:20 dan 70:30 untuk mengukur performa algoritma. Penelitian ini dengan kerangka pemikiran di Gambar 1. di bawah ini



Gambar 1. Kerangka Pemikiran

### 3. HASIL

Pengambilan dan pengumpulan Data Dataset sebanyak 2.000 berita berhasil diunduh dari [www.Kaggle.com](http://www.Kaggle.com) dan dikategorikan ke dalam lima kategori utama yaitu Politik, Ekonomi, Keamanan, Kesehatan, dan Kemiskinan. Dataset disimpan dalam format CSV untuk memudahkan pengolahan data.

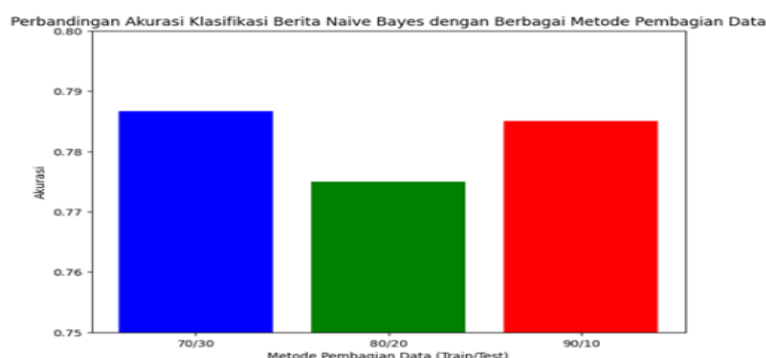
#### 3.1. Pembobotan Kata (TF-IDF)

Tahapan ini akan memberi bobot pada setiap kata untuk menghitung berapa kali sebuah kata muncul dalam setiap dokumen disebut Term Frequency (TF):

Gambar 2. Hasil TF-IDF

### 3.2. Klasifikasi Naïve Bayes

Penelitian ini mengimplementasikan algoritma Naïve Bayes untuk klasifikasi berita dengan tiga skenario pembagian data, yakni 70%-30%, 80%-20%, dan 90%-10% untuk data latih dan data uji. Pada skenario 70%-30%, model mencapai akurasi 78,66% dengan performa baik pada kategori "Keamanan" dan "Politik" namun gagal mengenali kategori "Ekonomi", "Kemiskinan", dan "Kesehatan". Rata-rata berbobot menunjukkan precision 0.66 dan recall 0.79. Pada skenario 80%-20%, model mencapai akurasi 77,5% dengan hasil serupa, di mana precision berbobot 0.64 dan recall 0.78. Skenario 90%-10% menghasilkan akurasi 78,5% dengan precision berbobot 0.68 dan recall 0.79, menunjukkan pola performa yang sama. Ketidakseimbangan dalam pengenalan kategori tertentu menjadi tantangan utama, meskipun model menunjukkan performa yang konsisten pada kategori tertentu seperti "Keamanan" dan "Politik." Peningkatan diperlukan untuk meningkatkan kemampuan model mengenali kategori yang kurang terwakili. Berikut grafik perbandingannya di gambar 3 di bawah ini



Gambar 3. Grafik Perbandingan 3 Skenario Pembagian Data

Gambar 3 Menunjukkan grafik masing-masing Skenario pembagian Data dengan 3 metode yaitu 70/30 dengan akurasi **78,66%** Klasifikasi Berita Dengan Data Latih 80% dan 20% data uji Model ini memiliki akurasi **77,50%**. Lalu untuk pembagian Data Latih 90% dan 10% data uji memiliki akurasi **78,50%**

### 3.3. Pengujian Confusion Matrix Naïve Bayes

Penelitian ini menguji model menggunakan 2000 data berita yang dibagi menjadi data latih dan data uji dengan tiga skenario: 70%-30%, 80%-20%, dan 90%-10%.

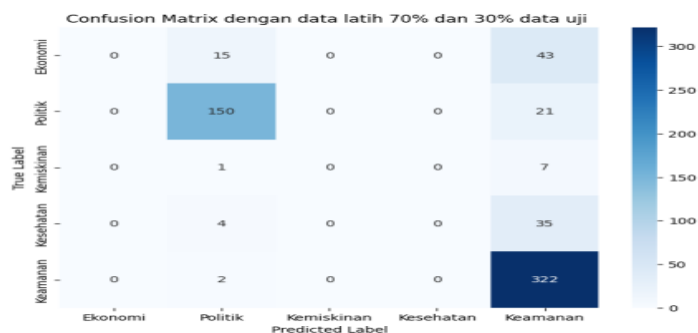
#### a. Pengujian terhadap data latih 70% dan 30% data uji

Berikut hasil pengujian *confusion matrix* untuk data latih 70% dan 30% data uji.

Tabel 1 Hasil *Confusion Matrix* pada data latih 70% dan 30% data uji

<i>Predicted Classification</i>	<i>Actual Classification</i>				
	Ekonomi Positive	Politik Positive	Kemiskinan Positive	Kesehatan Positive	Keamanan Positive
Ekonomi Negative	0	15	0	0	43
Politik Negative	0	150	0	0	21
Kemiskinan Negative	0	1	0	0	7
Kesehatan Negative	0	4	0	0	35
Keamanan Negative	0	2	0	0	322

Tabel 1 menghasilkan confusion matrix dari pengujian model dengan 70% data latih dan 30% data uji menunjukkan evaluasi performa klasifikasi untuk lima topik utama: Ekonomi, Politik, Kemiskinan, Kesehatan, dan Keamanan, dengan perbandingan antara prediksi model dan klasifikasi aktual.



Gambar 4 Hasil *Confusion Matrix* pada data latih 70% dan 30% data uji

Gambar 4 menunjukkan matriks kebingungan hasil pengujian model dengan data latih 70% dan data uji 30%, yang menggambarkan performa klasifikasi dalam lima kategori: Ekonomi, Politik, Kemiskinan, Kesehatan, dan Keamanan.

#### b. Pengujian terhadap data latih 80% dan 20% data uji

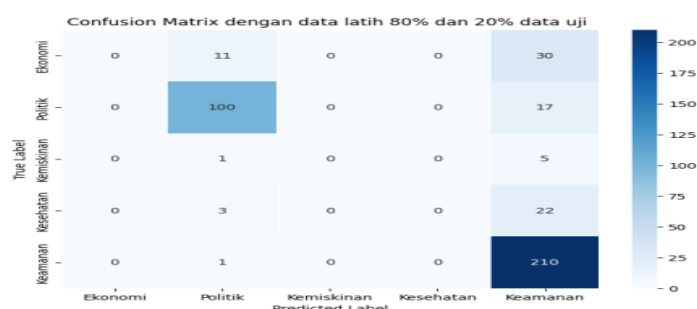
Berikut hasil pengujian *confusion matrix* untuk data latih 80% dan 20% data uji.

Tabel 2 Hasil *Confusion Matrix* pada data latih 80% dan 20% data uji

<i>Predicted Classification</i>	<i>Actual Classification</i>				
	Ekonomi Positive	Politik Positive	Kemiskinan Positive	Kesehatan Positive	Keamanan Positive
Ekonomi Negative	0	11	0	0	30

Politik Negative	0	100	0	0	17
Kemiskinan Negative	0	1	0	0	5
Kesehatan Negative	0	3	0	0	22
Keamanan Negative	0	1	0	0	210

Tabel 2 menampilkan confusion matrix dari pengujian model dengan 80% data latih dan 20% data uji, yang mengevaluasi kemampuan model mengidentifikasi lima kategori teks utama: Ekonomi, Politik, Kemiskinan, Kesehatan, dan Keamanan, dengan membandingkan hasil prediksi model dan klasifikasi aktual.



Gambar 5 Hasil Confusion Matrix pada data latih 80% dan 20% data uji

Gambar 5 menunjukkan confusion matrix hasil pengujian model dengan data latih 80% dan data uji 20%, yang menggambarkan performa klasifikasi dalam lima kategori: Ekonomi, Politik, Kemiskinan, Kesehatan, dan Keamanan.

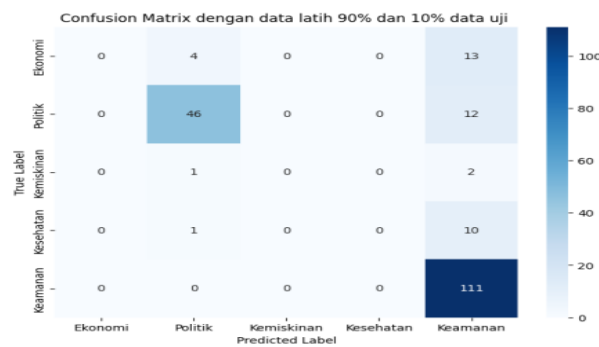
#### c. Pengujian terhadap data latih 90% dan 10% data uji

Berikut hasil pengujian *confusion matrix* untuk data latih 90% dan 10% data uji.

Tabel 3 Hasil *Confusion Matrix* pada data latih 90% dan 10% data uji

Predicted Classification	Actual Classification				
	Ekonomi Positive	Politik Positive	Kemiskinan Positive	Kesehatan Positive	Keamanan Positive
Ekonomi Negative	0	4	0	0	13
Politik Negative	0	46	0	0	12
Kemiskinan Negative	0	1	0	0	2
Kesehatan Negative	0	1	0	0	10
Keamanan Negative	0	0	0	0	111

Tabel 3 menunjukkan confusion matrix dari pengujian model klasifikasi dengan 90% data latih dan 10% data uji, yang mengevaluasi performa model dalam mengklasifikasikan lima topik utama: Ekonomi, Politik, Kemiskinan, Kesehatan, dan Keamanan, dengan membandingkan hasil prediksi model dan klasifikasi aktual.

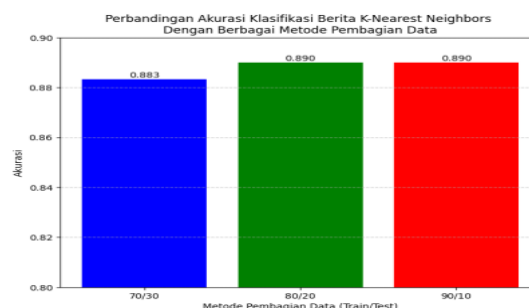


Gambar 6 Hasil *Confusion Matrix* pada data latih 90% dan 10% data uji

Gambar 6 menunjukkan matriks kebingungan dari model dengan data latih 90% dan data uji 10%, yang menggambarkan performa klasifikasi dalam lima kategori: Ekonomi, Politik, Kemiskinan, Kesehatan, dan Keamanan.

### 3.4. Klasifikasi K-Nearest Neighbors (KNN)

Penelitian ini menggunakan algoritma K-Nearest Neighbors (KNN) untuk klasifikasi berita setelah melalui proses praproses dan pembobotan kata (tf-idf). Model diuji pada tiga skenario pembagian data dengan nilai  $k=3$ . Pada skenario 70%-30%, model mencapai akurasi 88,33%, menunjukkan kinerja kuat pada kategori besar seperti "Keamanan" dan "Politik," namun perlu perbaikan untuk kategori kecil seperti "Kemiskinan." Skenario 80%-20% menghasilkan akurasi 85,25% dengan pola performa serupa, di mana rata-rata berbobot menunjukkan precision dan recall 0.85. Pada skenario 90%-10%, akurasi mencapai 85%, dengan kekuatan pada kategori besar seperti "Politik," tetapi "Kemiskinan" tidak terprediksi sama sekali. Secara keseluruhan, meskipun model bekerja baik pada kategori utama, peningkatan akurasi diperlukan untuk kategori dengan jumlah sampel lebih kecil, misalnya melalui pendekatan seperti cosine similarity.



Gambar 7. Grafik Perbandingan 3 Skenario Pembagian Data

Gambar 7 menunjukkan grafik akurasi dari tiga skenario pembagian data: 70%-30% dengan akurasi 87,3%, 80%-20% dengan akurasi 89,0%, dan 90%-10% juga dengan akurasi 89,0%.

### 3.5. Pengujian Confusion K-Nearest Neighbors (KNN)

Pengujian menggunakan confusion matrix dilakukan pada 2.000 data berita dengan tiga skenario pembagian data: 70%-30%, 80%-20%, dan 90%-10% untuk data latih dan data uji, menggunakan nilai  $k = 3$ .

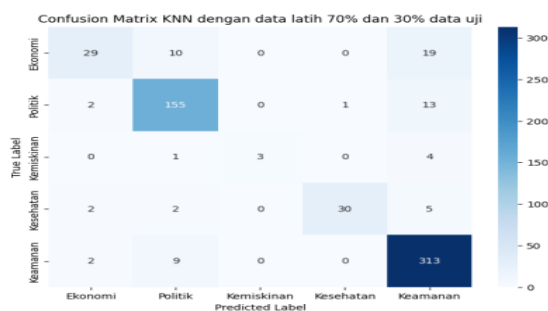
#### a. Pengujian terhadap data latih 70% dan 30% data uji

Berikut hasil pengujian *confusion matrix* untuk data latih 70% dan 30% data uji.

Tabel 4. Hasil Confusion Matrix pada data latih 70% dan 30% data uji KNN

Predicted Classification	Actual Classification				
	Ekonomi Positive	Politik Positive	Kemiskinan Positive	Kesehatan Positive	Keamanan Positive
Ekonomi Negative	29	10	0	0	19
Politik Negative	2	155	0	1	13
Kemiskinan Negative	0	1	3	0	4
Kesehatan Negative	2	2	0	30	5
Keamanan Negative	2	9	0	0	313

Tabel 4 menampilkan confusion matrix dari pengujian model klasifikasi dengan 70% data latih dan 30% data uji, yang digunakan untuk mengevaluasi performa model dalam mengklasifikasikan teks ke lima kategori utama: Ekonomi, Politik, Kemiskinan, Kesehatan, dan Keamanan. Matriks ini membandingkan hasil prediksi model dengan klasifikasi aktual dari data uji.



Gambar 8. Hasil Confusion Matrix pada data latih 70% dan 30% data uji KNN

Gambar 8 menampilkan matriks kebingungan hasil pengujian model dengan pembagian data 70% untuk pelatihan dan 30% untuk pengujian, yang menggambarkan kemampuan model dalam mengklasifikasikan data ke lima kategori: Ekonomi, Politik, Kemiskinan, Kesehatan, dan Keamanan.



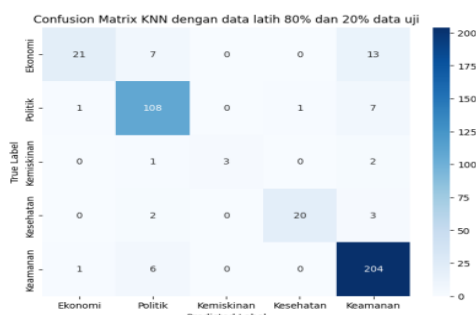
b. Pengujian terhadap data latih 80% dan 20% data uji

Berikut hasil pengujian *confusion matrix* untuk data latih 80% dan 20% data uji.

Tabel 5 Hasil Confusion Matrix pada data latih 80% dan 20% data uji KNN

<i>Predicted Classification</i>	<i>Actual Classification</i>				
	Ekonomi Positive	Politik Positive	Kemiskinan Positive	Kesehatan Positive	Keamanan Positive
Ekonomi Negative	21	7	0	0	13
Politik Negative	1	108	0	1	7
Kemiskinan Negative	0	1	3	0	2
Kesehatan Negative	0	2	0	20	3
Keamanan Negative	1	6	0	0	204

Tabel 5 menampilkan confusion matrix dari pengujian model klasifikasi dengan 80% data latih dan 20% data uji, yang digunakan untuk mengevaluasi performa model dalam mengidentifikasi lima topik utama: Ekonomi, Politik, Kemiskinan, Kesehatan, dan Keamanan. Matriks ini membandingkan hasil prediksi model dengan klasifikasi aktual dari data uji.



Gambar 9. Hasil Confusion Matrix pada data latih 80% dan 20% data uji KNN

Gambar 9 menampilkan matriks kebingungan dari model yang diuji dengan data latih 80% dan data uji 20%, yang menunjukkan performa klasifikasi model dalam mengelompokkan data ke lima kategori: Ekonomi, Politik, Kemiskinan, Kesehatan, dan Keamanan.

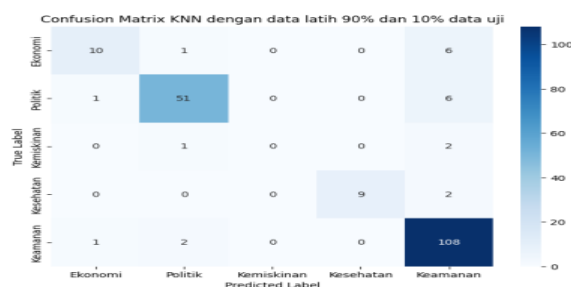
c. Pengujian terhadap data latih 90% dan 10% data uji

Berikut hasil pengujian *confusion matrix* untuk data latih 90% dan 10% data uji.

Tabel 6. Hasil *Confusion Matrix* pada data latih 90% dan 10% data uji KNN

<i>Predicted Classification</i>	<i>Actual Classification</i>				
	Ekonomi Positive	Politik Positive	Kemiskinan Positive	Kesehatan Positive	Keamanan Positive
Ekonomi Negative	10	1	0	0	6
Politik Negative	1	51	0	0	6
Kemiskinan Negative	0	1	0	0	2
Kesehatan Negative	0	0	0	9	2
Keamanan Negative	0	6	1	2	102

Tabel 6 menampilkan confusion matrix dari pengujian model dengan 90% data latih dan 10% data uji, yang mengevaluasi performa klasifikasi teks berdasarkan lima topik utama: Ekonomi, Politik, Kemiskinan, Kesehatan, dan Keamanan, dengan membandingkan hasil prediksi model dan data aktual.



Gambar 10. Hasil *Confusion Matrix* pada data latih 90% dan 10% data uji KNN

Gambar 10 menunjukkan matriks kebingungan dari model dengan pembagian data 90% pelatihan dan 10% pengujian, yang menggambarkan performa klasifikasi dalam lima kategori: Ekonomi, Politik, Kemiskinan, Kesehatan, dan Keamanan.

Untuk hasil akurasi pengujian lengkap pada tiap pembagian data menggunakan 7 nilai K dapat dilihat melalui tabel 4.8 berikut ini.

Tabel 11 Hasil Uji Akurasi

Nilai K	Akurasi (%)		
	<i>Split Data 70%:30%</i>	<i>Split Data 80%:20%</i>	<i>Split Data 90%:10%</i>
3	87.50%	85.25%	85.00%
5	87.50%	86.75%	84.50%
7	88.16%	87.50%	85.50%
9	88.50%	88.50%	89.50%
11	89.00%	89.00%	87.50%
13	89.50%	89.75%	88.50%
15	89.00%	89.75%	89.50%

Penelitian ini membandingkan kinerja model K-Nearest Neighbors (KNN) dan Naïve Bayes dalam klasifikasi berita menggunakan dataset dari Kaggle. Model KNN menunjukkan akurasi tertinggi pada semua skenario, yaitu 87,3% (70%-30%), 89,0% (80%-20%), dan 89,0% (90%-10%). Sementara itu, model Naïve Bayes mencatat akurasi 78,66% (70%-30%), 77,50% (80%-20%), dan 78,5% (90%-10%). Hasil ini menyoroti KNN sebagai model yang lebih unggul untuk klasifikasi opini, sedangkan Naïve Bayes memerlukan penyesuaian lebih lanjut untuk meningkatkan kinerjanya. Hasil penelitian menunjukkan bahwa:

- **KNN** memberikan akurasi tertinggi sebesar **83,25%** pada skenario pembagian data 80% latih dan 20% uji.
- **Naive Bayes** mencapai akurasi maksimal sebesar **78,83%** pada skenario pembagian data 70% latih dan 30% uji.
- Perbandingan menunjukkan bahwa KNN unggul dalam akurasi tetapi membutuhkan waktu lebih lama dalam pengolahan data, sedangkan Naive Bayes lebih cepat tetapi sedikit kurang akurat.

#### 4. KESIMPULAN

Penelitian ini bertujuan mengembangkan sistem klasifikasi berita bahasa Indonesia ke dalam lima kategori utama menggunakan algoritma K-Nearest Neighbor (KNN) dan Naïve Bayes. Dari 2000 dataset, algoritma KNN mencapai akurasi terbaik sebesar 83,25% pada pembagian data latih 80% dan data uji 20%, unggul dalam menangani data dengan distribusi kategori seimbang. Naïve Bayes mencapai akurasi 78,66% pada skenario yang sama, namun terbatas pada kategori dengan data berfrekuensi rendah. Proses preprocessing data, seperti cleaning, tokenizing, normalisasi, stopword removal, stemming, dan pembobotan TF-IDF, berkontribusi signifikan terhadap peningkatan akurasi. Penelitian ini memberikan kontribusi pada pengembangan sistem klasifikasi berita dan dapat diaplikasikan untuk rekomendasi atau analisis tren berita otomatis.

#### 5. REFERENSI

- [1] D. Julianti, "STRATEGI KEBIJAKAN PENGUATAN PELAYANAN DENGAN APLIKASI BERBASIS TEKNOLOGI INFORMASI," vol. 2, 2024.
- [2] N. I. Widiastuti, E. Rainarli, and K. E. Dewi, "Peringkasan dan Support Vector Machine pada Klasifikasi Dokumen," *J. Infotel*, vol. 9, no. 4, p. 416, 2017, doi: 10.20895/infotel.v9i4.312.
- [3] R. Nanda, E. Haerani, S. K. Gusti, and S. Ramadhani, "Klasifikasi Berita Menggunakan Metode Support Vector Machine," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 5, no. 2, pp. 269–278, 2022, doi: 10.32672/jnkti.v5i2.4193.
- [4] Muhammad Rifki Bahrul Uloom, Basuki Rahmat, and Made Hanindia Prami Swari, "Implementasi Metode CNN Dan K-Nearest Neighbor Untuk Klasifikasi

- Tingkat Kematangan Tanaman Cabai Rawit,” *Modem J. Inform. dan Sains Teknol.*, vol. 2, no. 3, pp. 112–123, 2024, doi: 10.62951/modem.v2i3.131.
- [5] A. P. Wijaya and H. A. Santoso, “Naive Bayes Classification pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government Naïve Bayes Classification on Document Classification to Identify E-Government Content,” *J. Appl. Intell. Syst.*, vol. 1, no. 1, pp. 48–55, 2016.
- [6] I. P. Putri, “Analisis Performa Metode K- Nearest Neighbor (KNN) dan Crossvalidation pada Data Penyakit Cardiovascular,” *Indones. J. Data Sci.*, vol. 2, no. 1, pp. 21–28, 2021, doi: 10.33096/ijodas.v2i1.25.
- [7] R. N. Mauliza and Y. R. Sipayung, “Penerapan Text Mining Dalam Menganalisis Pendapat Masyarakat Terhadap Pemilu 2024 Pada Media Sosial X Menggunakan Metode Naive Bayes,” *Technomedia J.*, vol. 9, no. 1, pp. 1–16, 2024, doi: 10.33050/tmj.v9i1.2212.
- [8] Wartumi, R. Kurniawan, and A. Y. Wijaya, “Analisis Data Sentimen Ulasan Pengguna Aplikasi Shopee di Google Play Store dengan Klasifikasi Algoritma Naïve Bayes,” *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 6, no. 1, pp. 164–170, 2024.
- [9] R. Yunitarini *et al.*, “KLASIFIKASI JAMU TRADISIONAL MADURA MENGGUNAKAN METODE K-NEAREST NEIGHBORS ( KNN ) DAN TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY ( TF-IDF ),” pp. 99–106, 2022.
- [10] H. S. Anggraheni, M. J. Naufal, and N. Yudistira, “DETEKSI SPAM BERBAHASA INDONESIA BERBASIS TEKS MENGGUNAKAN MODEL BERT TEXT-BASED INDONESIAN SPAM DETECTION USING THE BERT MODEL,” vol. 11, no. 6, pp. 1291–1301, 2024, doi: 10.25126/jtiik.2024118121.