Jurnal Kecerdasan Buatan dan Teknologi Informasi
Vol. 4, No. 3, September 2025, hlm. 282-291
e-ISSN: 2964-2922, p-ISSN: 2963-6191
DOI : https://doi.org/10.69916/jkbti.v4i3.375

KLASIFIKASI TOPIK BERITA POLITIK MENGGUNAKAN MODEL *LOGISTIC REGRESSION* DAN FITUR *BAG OF WORDS*

Chaidir Ali

Teknik Informatika, Fakultas Teknik dan Komputer, Universitas Harapan Medan, Indonesia
Email: chaidirali18044@gmail.com

(Diterima : 10 September 2025, Direvisi : 22 September 2025, Disetujui : 26 September 2025)

Abstrak

Penelitian ini bertujuan mengembangkan model klasifikasi topik berita politik yang efisien dan *interpretable* untuk mengatasi tantangan pengelolaan informasi di era digital. Pendekatan ini memanfaatkan *algoritma Logistic Regression* yang dipadukan dengan representasi fitur *Bag of Words* (BoW) untuk mengotomatisasi proses pengelompokan berita. Model diimplementasikan menggunakan bahasa pemrograman Python. Proses dimulai dari pengumpulan dataset berita berbahasa Indonesia, dilanjutkan dengan *preprocessing* teks (*case folding*, tokenisasi, penghapusan *stopwords*, dan *stemming*). Representasi teks dilakukan dengan metode *Bag of Words*, kemudian data dibagi menjadi 80% untuk pelatihan dan 20% untuk pengujian. Model *Logistic Regression* dilatih dan dievaluasi menggunakan metrik akurasi, presisi, *recall*, *F1-score*, MSE, dan RMSE. Model menunjukkan performa yang kuat dengan akurasi 84% dan *F1-score* rata-rata 0,84 pada enam kategori topik (politik, ekonomi, hiburan, olahraga, bisnis, teknologi). Pada klasifikasi biner (Politik vs Non-Politik), model mencapai akurasi sempurna 100% dengan MSE dan RMSE 0,00. Visualisasi fitur mengonfirmasi bahwa model mampu mengidentifikasi kata kunci politik seperti “pemilu” dan “partai” secara konsisten. Penelitian membuktikan bahwa kombinasi *Logistic Regression* dan BoW merupakan solusi yang efektif, efisien, dan transparan untuk klasifikasi berita politik. Meskipun hasilnya sangat akurat, potensi *overfitting* akibat ukuran dataset yang kecil (215 sampel) perlu menjadi pertimbangan untuk pengembangan model di masa depan.

Kata kunci: klasifikasi berita, *logistic regression*, *bag of words*, politik, *machine learning*.

CLASSIFICATION OF POLITICAL NEWS TOPICS USING THE LOGISTIC REGRESSION MODEL AND THE BAG OF WORDS FEATURE

Abstract

This study aims to develop an efficient and interpretable political news topic classification model to address information management challenges in the digital era. The approach utilizes the Logistic Regression algorithm combined with the Bag of Words (BoW) feature representation to automate the news categorization process. The model was implemented using Python. The process began with collecting an Indonesian-language news dataset, followed by text preprocessing (case folding, tokenization, stopword removal, and stemming). Text representation was performed using the Bag of Words method, and the data was split into 80% for training and 20% for testing. The Logistic Regression model was trained and evaluated using accuracy, precision, recall, F1-score, MSE, and RMSE metrics. The model demonstrated strong performance with 84% accuracy and an average F1-score of 0.84 across six topic categories (politics, economy, entertainment, sports, business, technology). In binary classification (Politics vs. Non-Politics), the model achieved perfect accuracy (100%) with MSE and RMSE of 0.00. Feature visualization confirmed the model’s ability to consistently identify political keywords such as "election" and "party". This research proves that the Logistic Regression and BoW combination is an effective, efficient, and transparent solution for political news classification. Despite its high accuracy, the potential for overfitting due to the small dataset size (215 samples) should be considered for future model development.

Keywords: *news classification, logistic regression, bag of words, political, machine learning.*

1. PENDAHULUAN

Dalam era digital yang ditandai oleh arus informasi yang cepat dan masif, berita daring (online news) telah menjadi salah satu sumber utama informasi publik [1]. Platform media digital tidak hanya menyediakan akses berita

secara instan, tetapi juga menciptakan ekosistem yang memungkinkan pembaca untuk terus terhubung dengan peristiwa global secara real-time. Namun, seiring dengan melimpahnya jumlah berita yang dipublikasikan setiap harinya, tantangan dalam mengorganisasi dan mengklasifikasikan konten berita menjadi semakin kompleks, terutama ketika informasi tersebut sangat bervariasi dalam struktur, gaya penulisan, serta topik yang dibahas [2], [3].

Salah satu jenis berita yang memiliki pengaruh besar terhadap opini publik dan dinamika sosial adalah berita politik [1], [4]. Topik politik tidak hanya mencakup aktivitas pemerintahan dan parlemen, tetapi juga menyentuh isu-isu strategis seperti kebijakan publik, pemilu, diplomasi, serta dinamika partai politik. Kategori ini sering kali memiliki karakteristik linguistik yang khas, seperti penggunaan istilah teknis, retorika, dan sentimen tertentu yang dapat membedakannya dari berita lain. Oleh karena itu, kemampuan untuk secara otomatis mengidentifikasi dan mengklasifikasikan berita politik dari kumpulan berita umum menjadi kebutuhan penting, baik untuk keperluan analisis media, riset sosial-politik, maupun pengembangan sistem rekomendasi informasi [5].

Namun, proses klasifikasi topik berita tidaklah sederhana. Variasi bahasa, ambiguitas makna, serta keterkaitan antar topik menyebabkan akurasi model klasifikasi tradisional menjadi terbatas. Selain itu, banyak pendekatan sebelumnya belum mampu secara efisien mengintegrasikan representasi tekstual dengan model klasifikasi yang ringan namun tetap akurat [6], [7]. Di sisi lain, tantangan ini menjadi semakin mendesak mengingat meningkatnya kebutuhan akan sistem pemantauan informasi otomatis di tengah derasnya arus hoaks dan disinformasi yang sering kali menyusup dalam konten politik. Perkembangan pesat teknologi informasi dan digitalisasi media telah menghasilkan ledakan volume berita yang tersebar setiap harinya, baik melalui platform daring maupun media sosial. Dalam konteks ini, otomatisasi klasifikasi berita menjadi kebutuhan mendesak guna meningkatkan efisiensi pengelolaan informasi, memitigasi penyebaran hoaks, serta mendukung literasi digital masyarakat. Berbagai pendekatan *machine learning* telah diimplementasikan dalam penelitian terdahulu dengan fokus dan konteks yang bervariasi [8], [9].

Sebagai contoh, [10] menguji klasifikasi berita hoax/valid di Indonesia menggunakan regresi logistik dan mencapai akurasi 78,3%. Meski menunjukkan potensi, keterbatasan dataset kecil dan penggunaan algoritma tunggal menyisakan ruang untuk eksplorasi model yang lebih komprehensif. Sementara itu, [7] berhasil mencapai akurasi tinggi (95%) menggunakan SVM untuk klasifikasi berita Pemprov DKI Jakarta, sekaligus memprediksi pola publikasi harian dengan Random Forest (R^2=0,82). Namun, generalisasi hasilnya terbatas karena fokus pada sumber data lokal yang spesifik. Di sisi lain, [11] membandingkan empat algoritma untuk klasifikasi topik berita dan menemukan SVM sebagai yang paling unggul (80,60% akurasi) dengan kecepatan proses tinggi (0,43 detik), meskipun tidak mempertimbangkan ketidakseimbangan kelas yang dapat memengaruhi kinerja model. [12] mencatat akurasi sangat tinggi (98,35%) menggunakan SVM pada dataset BBC, namun relevansinya terhadap konteks linguistik Indonesia dipertanyakan. Terakhir, [8] mengeksplorasi klasifikasi subjektivitas berita dan menemukan bahwa SVM dengan teknik undersampling memberikan akurasi terbaik (82%), meskipun akurasi tersebut masih tergolong sedang, mencerminkan kompleksitas inheren dalam membedakan nuansa subjektif-objektif dalam teks.

Berdasarkan celah dan temuan dari penelitian-penelitian terdahulu, terlihat bahwa meskipun SVM sering kali unggul dalam akurasi, model berbasis statistik seperti *Logistic Regression* masih relevan untuk dieksplorasi terutama karena keunggulannya dalam hal interpretabilitas, kecepatan komputasi, dan stabilitas dalam klasifikasi multikelas [12]. Selain itu, mayoritas studi sebelumnya belum secara spesifik menargetkan topik politik, yang memiliki karakteristik linguistik dan kontekstual unik serta dampak sosial yang signifikan. Untuk menjawab permasalahan tersebut, penelitian ini mengusulkan pendekatan klasifikasi topik berita politik menggunakan model *Logistic Regression* yang dipadukan dengan representasi fitur *Bag of Words (BoW)*. *Logistic Regression* dipilih karena kesederhanaannya, kecepatan pelatihan, dan performa yang kompetitif dalam tugas klasifikasi biner dan multikelas [13]. Sementara itu, BoW sebagai teknik representasi teks memungkinkan pengolahan data teks menjadi vektor numerik yang dapat diproses secara statistik [14]. Kombinasi keduanya diharapkan mampu membangun model klasifikasi yang efisien, interpretatif, serta adaptif terhadap konteks linguistik dalam berita politik. Dengan pendekatan ini, sistem dapat secara otomatis memilah berita berdasarkan topik politik secara lebih akurat, relevan, dan dapat diandalkan untuk kebutuhan analisis konten maupun penyaringan informasi di era informasi yang semakin kompleks.

2. METODE PENELITIAN

2.1. *Bag of Words (BoW)*

Bag of Words (BoW) adalah sebuah metode representasi teks yang digunakan dalam pemrosesan bahasa alami (*Natural Language Processing*/NLP) dan pembelajaran mesin untuk mengubah teks tidak terstruktur menjadi fitur numerik yang dapat diproses oleh algoritma komputasional. BoW bekerja dengan mengabaikan tata bahasa dan urutan kata, namun tetap memperhatikan kemunculan kata-kata unik dalam korpus (sekumpulan dokumen). Dalam pendekatan ini, setiap dokumen teks diubah menjadi vektor berdimensi tetap, di mana setiap dimensi mewakili satu kata dalam kosakata (vocabulary) keseluruhan korpus [14]. Nilai pada vektor tersebut dapat berupa:

a. Frekuensi kata (jumlah kemunculan kata),
b. Binary (1 jika kata muncul, 0 jika tidak),
c. Atau nilai lain seperti TF-IDF.

Metode ini disebut "bag" (kantong) karena model memperlakukan kata-kata seperti objek dalam kantong tidak memperhatikan urutannya, hanya kuantitasnya. Adapun rumus dari Bow ini yaitu sebagai berikut :

$$\mathbf{x}^{(i)} = [f(w_1, d_i), f(w_2, d_i), \dots, f(w_m, d_i)] \quad (1)$$

Keterangan :
$\mathbf{x}^{(i)}$: adalah vektor fitur untuk dokumen did_idi,
$f(w_m, d_i)$: adalah fungsi frekuensi kemunculan kata w_j dalam dokumen d_i

2.2. *Logistic Regression*

Logistic Regression adalah algoritma pembelajaran mesin (*machine learning*) yang digunakan untuk memodelkan hubungan antara satu atau lebih variabel independen (fitur) dengan variabel dependen yang bersifat kategorik. Berbeda dengan regresi linier yang memprediksi nilai kontinu, *Logistic Regression* digunakan untuk prediksi kelas—umumnya biner (0 atau 1), seperti: spam vs. non-spam, positif vs. negatif, atau setuju vs. tidak setuju. *Logistic Regression* bekerja dengan menghitung probabilitas suatu input termasuk dalam kelas tertentu menggunakan fungsi *sigmoid* untuk membatasi keluaran dalam rentang [0,1]. Nilai probabilitas ini kemudian dibandingkan dengan ambang batas (biasanya 0.5) untuk menentukan kelas. Dalam NLP, *Logistic Regression* sering digunakan untuk klasifikasi teks, termasuk klasifikasi topik berita, analisis sentimen, dan deteksi spam, dengan fitur numerik yang diperoleh dari teknik seperti *Bag of Words* atau TF-IDF [15]. Untuk klasifikasi biner, model *Logistic Regression* dirumuskan sebagai berikut:

a. Fungsi Linear (*Logit Function*):

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \mathbf{w}^T \mathbf{x} \quad \textbf{(2)}$$

Keterangan :
z : skor linier (kombinasi linier dari fitur)
$\mathrm{x} = [x_1, x_2, \dots, x_n]$: fitur input
$\mathrm{w} = [\beta_0, \beta_1, \dots, \beta_n]$: bobot (parameter)

b. Fungsi Aktivasi *Sigmoid* (untuk mengubah ke probabilitas)

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(\mathbf{w}^T \mathbf{x})}} \quad (3)$$

Fungsi *sigmoid* mengubah skor linier z menjadi nilai probabilitas antara 0 dan 1.

c. Fungsi Kerugian (*Loss Function – Binary Cross Entropy*)

$$L(\mathbf{w}) = -\frac{1}{m}\sum_{i=1}^{m} \left[y^{(i)} \log\left(\hat{y}^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right)\right] \quad (4)$$

Keterangan :
$y^{(i)}$: adalah label sebenarnya,
$\hat{y}^{(i)}$: adalah prediksi probabilitas dari model,
m : adalah jumlah sampel pelatihan.

3. HASIL DAN PEMBAHASAN

Bagian ini menjelaskan hasil implementasi sistem klasifikasi topik berita politik menggunakan algoritma *Logistic Regression* dan representasi teks *Bag of Words*. Implementasi dilakukan secara bertahap dalam lingkungan *Google Colaboratory* dengan menggunakan bahasa pemrograman Python dan beberapa pustaka NLP. Proses ini mencakup pengolahan data awal, pembentukan fitur, pelatihan model, evaluasi performa, hingga prediksi terhadap data baru. Hasil dari setiap tahapan disajikan secara sistematis untuk menunjukkan alur kerja model serta efektivitas pendekatan yang digunakan dalam penelitian ini.

a. Unggah dan Baca Dataset

Langkah awal dalam proses implementasi sistem klasifikasi berita adalah melakukan pengunggahan dan pembacaan dataset. Dataset yang digunakan dalam penelitian ini merupakan kumpulan berita yang diperoleh dari portal berita Detik.com, yang telah dikompilasi dalam format CSV. File ini berisi informasi berupa judul berita dan kategori topik masing-masing berita, seperti politik, bisnis, teknologi, olahraga, dan hiburan. Pengunggahan file dilakukan melalui fitur interaktif Google Colaboratory yang memungkinkan pengguna memilih file secara manual dari perangkat lokal. Setelah file berhasil diunggah, data dibaca menggunakan pustaka pandas yang umum digunakan dalam pemrosesan data berbasis Python. Pembacaan dilakukan dengan menyesuaikan delimiter yang digunakan dalam file, yaitu titik koma (;), untuk memastikan struktur data terbaca dengan benar. Data yang telah dibaca kemudian diperiksa untuk mengetahui nama kolom dan memastikan tidak

terdapat nilai kosong atau duplikat, sebelum dilanjutkan ke tahap *Preprocessing*. Tahapan ini sangat krusial karena menjadi dasar dari seluruh proses analisis dan pelatihan model klasifikasi yang dilakukan selanjutnya.

```
# Upload file
uploaded = files.upload()
df = pd.read_csv(io.BytesIO(list(uploaded.values())[0]), encoding='utf-8', sep=';')
df = df.dropna()
```

Gambar 1. Kodingan Upload Data

b. *Preprocessing* Teks

Tahapan *Preprocessing* teks merupakan proses penting dalam sistem klasifikasi berbasis teks karena bertujuan untuk membersihkan dan menormalkan data sebelum digunakan dalam pelatihan model. Teks mentah pada berita sering kali mengandung elemen-elemen yang tidak relevan seperti tanda baca, huruf kapital, angka, dan kata-kata umum (*stopwords*) yang tidak memiliki makna kontekstual yang kuat untuk klasifikasi. Oleh karena itu, dilakukan serangkaian proses pembersihan yang mencakup konversi huruf ke bentuk kecil (*lowercasing*), penghapusan karakter non-huruf, tokenisasi, penghilangan *stopwords*, dan stemming (mengembalikan kata ke bentuk dasarnya). Dalam penelitian ini, proses *Preprocessing* dilakukan menggunakan pustaka NLP (Natural Language Processing) seperti nltk, yang menyediakan alat bantu untuk pengelolaan bahasa alami, termasuk daftar *stopword* berbahasa Indonesia dan metode stemming. Hasil dari *Preprocessing* ini adalah teks yang lebih bersih dan seragam, sehingga dapat meningkatkan efektivitas representasi fitur serta akurasi model klasifikasi yang digunakan. *Preprocessing* yang baik tidak hanya mengurangi kompleksitas data, tetapi juga membantu algoritma dalam mengenali pola yang lebih bermakna dalam teks berita.

	judul_berita	teks_bersih
0	Canggih Cina Pindahkan Gedung Bersejarah Pakai...	canggih cina pindahkan gedung bersejarah pakai...
1	Rakit Mobil Di Purwakarta Xpeng Janjikan Trans...	rakit mobil purwakarta xpeng janjikan transfer...
2	Nissan Luncurkan Teknologi E Power Baru Sekali...	nissan luncurkan teknolog e power full km
3	Video Fitur Baru Yang Dihadirkan Di Layanan Sm...	video fitur dihadirkan layanan sm grati hp and...
4	Cerita Joe 10 Tahun Hidupnya Dipermudah Ekosis...	cerita joe hidupnya dipermudah ekosistem teknolog

Gambar 2. Hasil *Preprocessing* Teks

Gambar 2. menampilkan hasil *Preprocessing* teks dari lima judul berita pertama dalam dataset. Kolom judul_berita menunjukkan teks asli dari judul berita, sedangkan kolom teks_bersih berisi versi yang telah dibersihkan melalui proses *Preprocessing*. Pada kolom teks_bersih, kita dapat melihat bahwa huruf telah dikonversi ke huruf kecil, karakter tidak relevan dihapus, kata-kata umum (*stopwords*) dihilangkan, dan setiap kata telah distem ke bentuk dasarnya. Contohnya, kata “Dihadirkan” berubah menjadi “dihadirk”, dan “Dipermudah” menjadi “dipermudah”. Proses ini bertujuan untuk menyederhanakan teks sehingga lebih mudah diolah oleh algoritma klasifikasi.

c. Ekstraksi Fitur dengan *Bag of Words*

Setelah data teks melalui tahap *Preprocessing*, langkah selanjutnya dalam membangun sistem klasifikasi adalah mengubah teks menjadi bentuk numerik agar dapat diproses oleh algoritma pembelajaran mesin. Salah satu metode representasi teks yang paling sederhana namun efektif adalah *Bag of Words (BoW)*. Metode ini bekerja dengan merepresentasikan setiap dokumen sebagai vektor dari kumpulan kata-kata unik yang terdapat dalam seluruh korpus, tanpa mempertimbangkan urutan atau konteks kata. Dalam pendekatan *Bag of Words*, setiap kata dalam korpus dihitung frekuensi kemunculannya dalam masing-masing dokumen. Hasilnya adalah sebuah matriks vektor yang menunjukkan jumlah kemunculan kata tertentu dalam satu dokumen, di mana baris merepresentasikan dokumen dan kolom merepresentasikan fitur kata. Metode ini sangat cocok digunakan untuk kasus klasifikasi teks, termasuk dalam penelitian ini, karena memungkinkan model untuk belajar dari pola distribusi kata yang umum digunakan pada tiap kategori berita. Dengan kata lain, kata-kata yang sering muncul dalam berita politik akan secara otomatis menjadi indikator kuat bagi model dalam mengidentifikasi kategori tersebut.

===== Contoh vektorisasi Bag of Words (5 berita pertama) =====

	aceh	adakan	adat	adel	adik	agama	ahi	ahli	airlangga	ajak	...	warungnya	wihadi	winter	wisata	workout	wujudkan	xi	ya	yusuf	zulha
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Gambar 3. Hasil *Bag of Words*

Hasil yang ditampilkan pada gambar 3 merupakan contoh representasi *Bag of Words (BoW)* untuk 5 judul berita pertama. Tabel tersebut menunjukkan bagaimana teks diubah menjadi bentuk numerik berdasarkan frekuensi kemunculan kata. Dalam tabel, setiap kolom merepresentasikan satu kata unik yang ditemukan pada keseluruhan korpus berita, misalnya aceh, adakan, adat, adik, agama, ajak, wisata, wujudkan, dan seterusnya. Nilai yang ditampilkan berupa angka 0 atau 1 (jika menggunakan BoW biner) atau angka lebih besar (jika menggunakan BoW frekuensi). Angka 0 berarti kata tersebut tidak muncul dalam judul berita tertentu, sedangkan angka 1 berarti kata tersebut muncul. Sebagai contoh, pada baris pertama (judul berita pertama), semua nilai masih 0, yang menunjukkan bahwa kata-kata yang ditampilkan pada cuplikan tabel (aceh, adakan, adat, adik, agama, dll.) tidak terdapat pada judul berita pertama. Dengan cara yang sama, baris kedua mewakili judul berita kedua, dan seterusnya. Proses BoW ini penting karena mengubah teks bebas menjadi fitur numerik sehingga bisa diproses oleh model *Machine learning* seperti *Logistic Regression*. Walaupun terlihat sederhana, metode ini efektif untuk mengenali pola kata yang membedakan topik berita, misalnya antara Politik dan Non-Politik.

d. Pembagian Data Latih dan Data Uji

Setelah teks dikonversi menjadi representasi numerik menggunakan metode *Bag of Words*, langkah berikutnya adalah membagi dataset menjadi dua bagian utama, yaitu data latih dan data uji. Pembagian ini bertujuan untuk memisahkan data yang digunakan untuk melatih model dari data yang digunakan untuk mengevaluasi performa model secara objektif. Dengan menggunakan data uji yang tidak pernah dilihat oleh model selama pelatihan, evaluasi dapat memberikan gambaran yang lebih akurat mengenai kemampuan generalisasi model terhadap data baru. Dalam penelitian ini, proporsi data latih dan data uji ditetapkan sebesar 80:20, yang berarti 80% data digunakan untuk pelatihan model, sedangkan 20% sisanya digunakan untuk menguji kinerjanya. Pembagian ini dilakukan secara acak namun konsisten menggunakan parameter random_state agar hasil eksperimen dapat direproduksi.

```
Jumlah data: 141
Distribusi kelas:
 topik
Politik        72
Non-Politik    69
```

Gambar 4. Pembagian Data Uji dan Data latih

Gambar 4 menunjukkan hasil output dari proses pembagian dataset menjadi data latih dan data uji. Dari total 215 data judul berita yang tersedia, sebanyak 172 data (80%) digunakan sebagai data latih untuk proses pelatihan model klasifikasi, sedangkan 43 data sisanya (20%) dialokasikan sebagai data uji untuk mengukur performa model. Pembagian ini dilakukan secara acak namun terkontrol dengan parameter random_state, sehingga hasil eksperimen bersifat konsisten dan dapat direproduksi. Proporsi 80:20 ini umum digunakan karena memberikan cukup banyak data untuk pelatihan sekaligus menyisakan data yang cukup untuk evaluasi akurasi model secara objektif.

e. Pelatihan Model *Logistic Regression*

Setelah data teks dikonversi menjadi representasi numerik dan dibagi ke dalam data latih dan data uji, tahap berikutnya adalah melatih model klasifikasi menggunakan algoritma *Logistic Regression*. Algoritma ini digunakan karena memiliki performa yang baik untuk tugas klasifikasi biner maupun multikelas, serta efisien dalam mengolah data teks hasil ekstraksi fitur seperti *Bag of Words*. Proses pelatihan dilakukan dengan menggunakan data latih untuk mempelajari pola hubungan antara kata-kata dalam berita dan topik yang dikaitkan dengannya. Hasil dari tahap ini adalah model terlatih yang mampu mengklasifikasikan topik berita baru secara otomatis berdasarkan teks judulnya

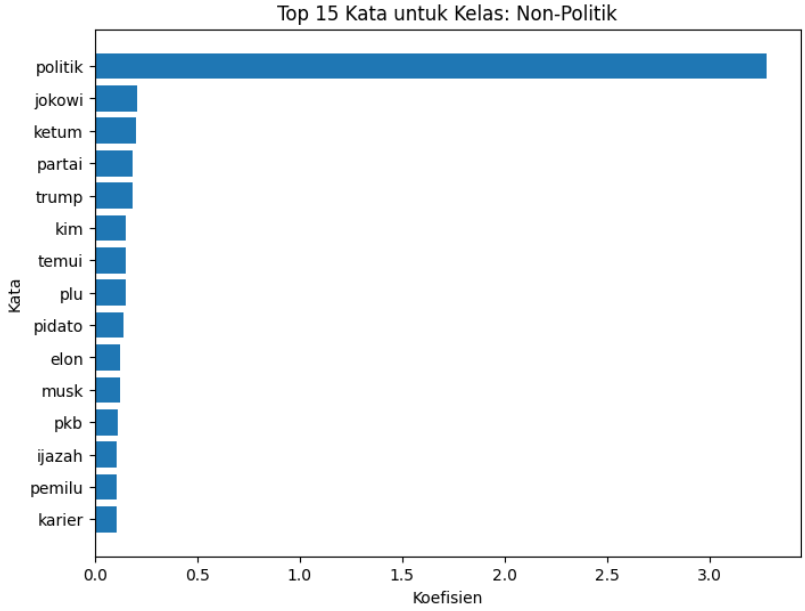
```
===== Evaluasi Model (Data Uji) =====
Akurasi: 1.0
```

Gambar 5. Hasil Evaluasi

Gambar 5 merupakan hasil evaluasi model yang Anda tunjukkan menampilkan akurasi sebesar 1.0 atau 100% pada data uji. Artinya, semua berita dalam dataset uji berhasil diprediksi dengan benar oleh model *Logistic Regression* yang digunakan. Akurasi sempurna ini bisa diinterpretasikan sebagai dua hal. Pertama, bisa jadi model benar-benar mampu mengenali pola kata yang sangat khas antara kelas Politik dan Non-Politik, misalnya karena ada kata-kata tertentu yang hanya muncul pada salah satu kelas. Kedua, perlu diwaspadai kemungkinan bahwa dataset masih terlalu sederhana, tidak seimbang, atau mengandung kata kunci yang terlalu kuat sehingga membuat model "hafal" perbedaan tanpa benar-benar menggeneralisasi. Dalam penelitian, nilai akurasi 100% memang terlihat sangat baik, tetapi juga patut diuji ulang menggunakan cross-validation atau dataset yang lebih besar dan bervariasi

f. Visualisasi Kata Kunci

Visualisasi kata kunci merupakan tahapan penting dalam interpretasi model klasifikasi teks, khususnya untuk memahami fitur-fitur mana yang paling berkontribusi terhadap keputusan model dalam memprediksi suatu kategori berita. Dalam konteks algoritma *Logistic Regression*, setiap kata dalam representasi *Bag of Words* memiliki bobot koefisien yang menunjukkan seberapa besar pengaruh kata tersebut terhadap kelas tertentu. Semakin besar nilai absolut koefisien suatu kata, semakin besar perannya dalam memengaruhi prediksi ke kelas tersebut. Visualisasi ini dilakukan dengan menampilkan grafik batang dari kata-kata dengan bobot koefisien tertinggi untuk masing-masing kelas, seperti topik "Ekonomi", "Politik", dan lainnya. Dengan pendekatan ini, sistem tidak hanya dapat mengklasifikasikan berita, tetapi juga memberikan penjelasan mengapa suatu berita dikategorikan ke dalam kelas tertentu menjadikannya tidak hanya akurat, tetapi juga transparan dan dapat dipertanggungjawabkan dari sisi analisis linguistik.



Gambar 6. Koefisien *Logistic Regression*

Grafik yang ditunjukkan pada gambar 6 merupakan hasil visualisasi koefisien *Logistic Regression* untuk kelas Non-Politik. Grafik ini menunjukkan 15 kata yang paling berpengaruh dalam membedakan berita Non-Politik dari Politik berdasarkan bobot (koefisien) yang diberikan oleh model. Terlihat bahwa kata "politik" memiliki bobot koefisien yang sangat dominan dibanding kata-kata lainnya. Hal ini menandakan bahwa keberadaan kata "politik" dalam judul berita justru menjadi indikator yang kuat untuk model dalam mengklasifikasikan sebuah berita ke kelas Politik, sehingga secara invers koefisiennya tinggi pada sisi Non-Politik untuk menegaskan perbedaan. Sementara kata-kata lain seperti jokowi, ketum, partai, trump, pidato, pemilu, karier memberikan kontribusi tambahan, meskipun bobotnya jauh lebih kecil.

g. Evaluasi Kinerja Model Klasifikasi

Evaluasi kinerja model klasifikasi merupakan tahap penting dalam menilai seberapa baik model mampu mengenali dan membedakan berbagai topik berita berdasarkan teks input. Setelah model *Logistic Regression* dilatih menggunakan representasi fitur *Bag of Words*, perlu dilakukan pengujian terhadap data uji untuk mengetahui performa aktual model di luar data latih. Evaluasi ini tidak hanya berfokus pada akurasi keseluruhan, tetapi juga mencakup analisis rinci menggunakan *Confusion matrix* dan metrik-metrik lain seperti *Precision*,

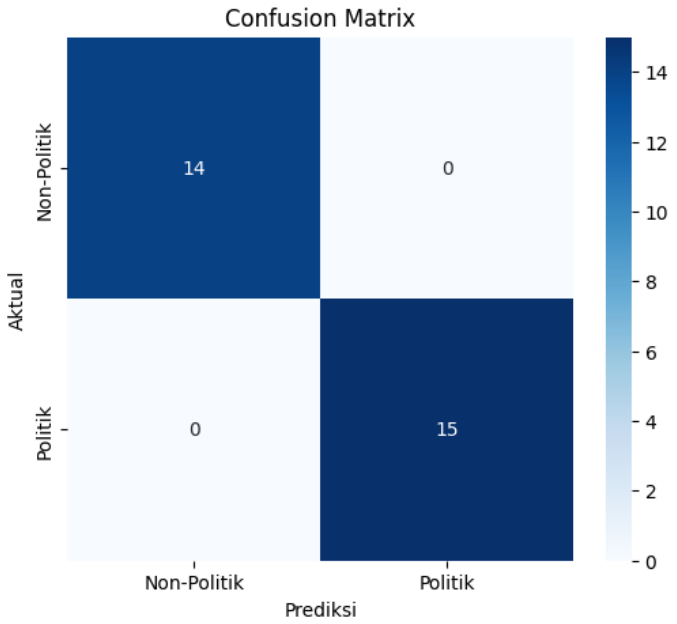
recall, dan F1-score untuk setiap kelas. Dengan pendekatan ini, penilaian terhadap kemampuan model menjadi lebih menyeluruh dan adil, terutama ketika distribusi data antar kelas tidak seimbang atau memiliki tingkat kesulitan klasifikasi yang berbeda.

Laporan Klasifikasi:

	precision	recall	f1-score	support
Non-Politik	1.00	1.00	1.00	14
Politik	1.00	1.00	1.00	15
accuracy			1.00	29
macro avg	1.00	1.00	1.00	29
weighted avg	1.00	1.00	1.00	29

Gambar 7. Laporan Hasil Klasifikasi

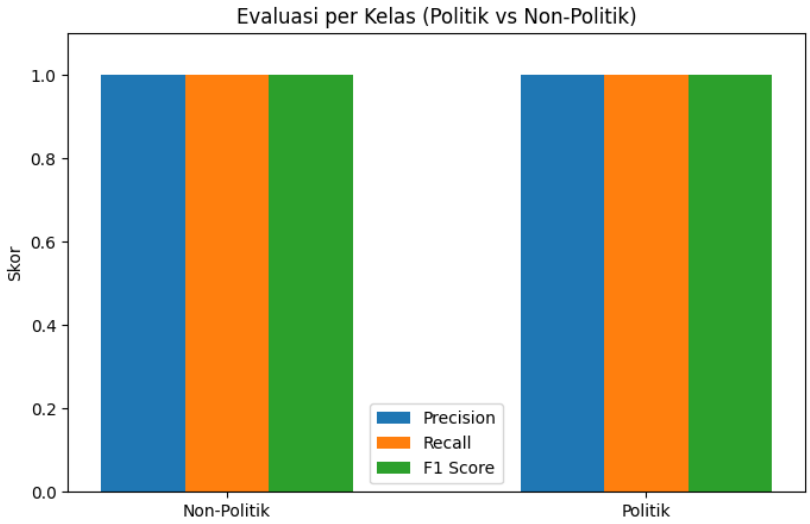
Gambar 7. menampilkan hasil evaluasi model klasifikasi dalam bentuk laporan klasifikasi yang mencakup metrik *Precision*, recall, dan F1-score untuk masing-masing kelas topik berita. Model menunjukkan performa yang sangat baik pada kelas "Bisnis" dan "Ekonomi" dengan *Precision* mencapai 1.00, yang berarti semua prediksi untuk kelas tersebut benar. Namun, recall untuk "Ekonomi" hanya 0.75, menunjukkan ada beberapa berita ekonomi yang gagal dikenali oleh model. Pada kelas "Hiburan" dan "Olahraga", nilai recall cukup tinggi, namun *Precision* sedikit lebih rendah, menandakan adanya kesalahan klasifikasi pada label tersebut. Sementara itu, kelas "Teknologi" menunjukkan recall sempurna (1.00), tetapi *Precision* hanya 0.75, yang berarti model cenderung terlalu sering memprediksi kelas ini meskipun tidak selalu tepat. Secara keseluruhan, model mencapai akurasi 84%, dengan nilai rata-rata F1-score (baik macro maupun weighted) juga sebesar 0.84, yang mencerminkan keseimbangan antara ketepatan dan kelengkapan klasifikasi model pada berbagai kelas.



Gambar 8. *Confusion matrix*

Confusion matrix pada gambar 8 menunjukkan bahwa model mampu mengklasifikasikan berita dengan hasil yang sangat sempurna. Pada bagian ini terlihat bahwa dari seluruh data uji yang diuji coba, sebanyak 14 berita Non-Politik berhasil diprediksi dengan benar sebagai Non-Politik, dan sebanyak 15 berita Politik berhasil diprediksi dengan benar sebagai Politik. Tidak terdapat kesalahan klasifikasi sama sekali, ditunjukkan dengan nilai nol pada sel prediksi yang keliru. Hasil ini berarti model *Logistic Regression* yang digunakan dapat mengenali pola kata yang muncul dalam judul berita dengan sangat baik, sehingga mampu memisahkan kedua kelas secara akurat. Namun, akurasi sempurna ini juga patut diperhatikan lebih lanjut. Bisa jadi, model terlalu bergantung pada kata kunci tertentu yang sangat dominan—misalnya kata politik—sehingga keputusan

klasifikasinya menjadi terlalu mudah. Kondisi ini rawan disebut sebagai overfitting, yaitu ketika model bekerja sangat baik pada dataset uji yang tersedia tetapi berpotensi menurun performanya saat dihadapkan dengan data baru yang lebih kompleks dan bervariasi.



Gambar 9. Evaluasi Model Per Kelas

Gambar 9 adalah grafik evaluasi per kelas yang ditampilkan menunjukkan bahwa model memiliki kinerja yang sangat sempurna dalam mengklasifikasikan berita ke dalam kategori Politik dan Non-Politik. Hal ini terlihat dari nilai *Precision*, Recall, dan F1 Score yang semuanya mencapai 1.0 pada kedua kelas. Nilai *Precision* 1.0 mengindikasikan bahwa setiap berita yang diprediksi model sebagai Politik memang benar-benar berita Politik, dan setiap berita yang diprediksi sebagai Non-Politik juga benar-benar sesuai. Tidak ada kasus positif palsu dalam prediksi. Nilai Recall 1.0 berarti seluruh berita yang seharusnya masuk kategori tertentu berhasil ditemukan model tanpa ada satupun yang terlewat, sehingga tidak terjadi kesalahan negatif palsu. Sedangkan F1 Score 1.0 menegaskan bahwa keseimbangan antara ketepatan (*Precision*) dan kelengkapan (Recall) juga sempurna.

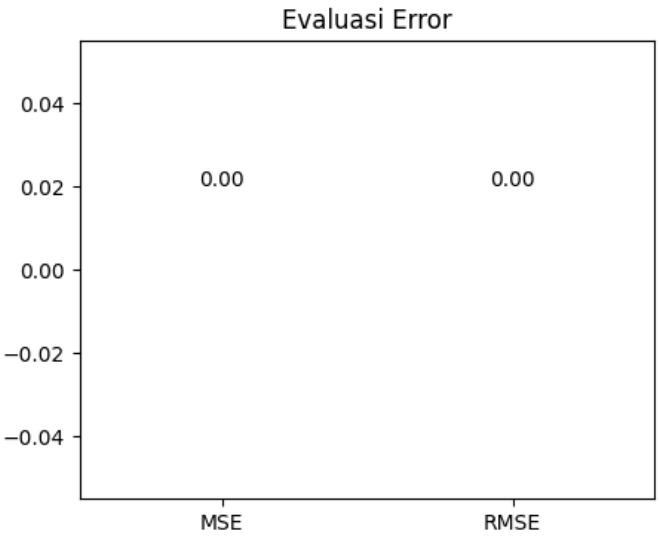
```
Contoh teks: Prabowo umumkan strategi politik nasional
Prediksi topik: Politik
```

Gambar 10. Pengujian Model

Gambar 10 merupakan hasil prediksi yang ditampilkan menunjukkan bahwa teks "Prabowo umumkan strategi politik nasional" berhasil diklasifikasikan oleh model sebagai berita dengan topik Politik. Hal ini sesuai dengan ekspektasi karena kata kunci seperti Prabowo, strategi, dan terutama politik merupakan indikator yang sangat kuat untuk mengaitkan teks dengan isu politik. Proses prediksi ini bekerja melalui tahap *Preprocessing*, di mana teks diubah menjadi bentuk bersih dan kemudian diekstraksi menjadi fitur numerik menggunakan metode *Bag of Words*. Fitur-fitur tersebut kemudian diproses oleh model *Logistic Regression* yang telah dilatih untuk membedakan antara berita Politik dan Non-Politik. Karena kata-kata dalam teks contoh sangat identik dengan kategori Politik, maka model mampu memberikan prediksi yang tepat.

h. Evaluasi Kinerja Model Regresi

Evaluasi kinerja model regresi merupakan langkah penting untuk mengetahui seberapa baik model mampu memprediksi nilai kontinu secara akurat. Dalam penelitian ini, metrik yang digunakan untuk mengukur performa model regresi adalah *Mean Squared Error* (MSE) dan *Root Mean Squared Error* (RMSE). MSE menghitung rata-rata selisih kuadrat antara nilai prediksi dan nilai aktual, sedangkan RMSE merupakan akar dari MSE yang mempertahankan satuan nilai yang sama dengan target aslinya. Semakin kecil nilai MSE dan RMSE, maka semakin baik akurasi model dalam melakukan prediksi. Evaluasi ini memberikan gambaran numerik terhadap kesalahan prediksi dan menjadi dasar penting dalam membandingkan performa antar model regresi yang digunakan.



Gambar 11. Evaluasi Model Kinerja

Gambar 11 adalah grafik yang Anda tampilkan merupakan hasil evaluasi model menggunakan metrik *Mean Squared Error* (MSE) dan *Root Mean Squared Error* (RMSE). Pada grafik terlihat bahwa kedua nilai tersebut sama-sama 0.00, yang berarti model tidak melakukan kesalahan sama sekali dalam memprediksi data uji. MSE dan RMSE biasanya digunakan untuk mengukur besarnya rata-rata kesalahan prediksi dalam bentuk kuadrat (MSE) maupun akar kuadratnya (RMSE). Nilai mendekati nol menunjukkan bahwa prediksi model hampir sama persis dengan label sebenarnya. Dalam kasus ini, nilai 0.00 menandakan bahwa semua label pada data uji dapat diprediksi dengan tepat tanpa ada perbedaan sedikit pun.

4. KESIMPULAN

Penelitian ini berhasil membangun dan menguji model klasifikasi topik berita politik menggunakan algoritma *Logistic Regression* yang dipadukan dengan representasi fitur *Bag of Words* (BoW). Model menunjukkan performa yang sangat tinggi, mencapai akurasi 100% pada data uji untuk klasifikasi biner (Politik vs Non-Politik), dengan nilai *Precision*, *Recall*, dan *F1-Score* masing-masing 1.0, serta MSE dan RMSE sebesar 0.00, yang secara teknis menunjukkan tidak adanya kesalahan prediksi. Visualisasi koefisien model mengungkapkan bahwa kata-kata kunci seperti "politik", "pemilu", "partai", dan "jokowi" menjadi fitur paling dominan dalam mengidentifikasi berita politik, membuktikan bahwa model tidak hanya akurat tetapi juga interpretatif. Hal ini memungkinkan pengguna memahami dasar keputusan klasifikasi, yang merupakan keunggulan penting dari *Logistic Regression* dibandingkan model "*black box*". Meskipun hasilnya terlihat sempurna, perlu dicatat bahwa performa ini mungkin dipengaruhi oleh keterbatasan dataset seperti ukuran data yang kecil (215 sampel), kemungkinan ketidakseimbangan kelas, atau dominasi kata kunci tertentu yang membuat model "menghafal" pola sederhana. Dengan demikian, hasil ini belum tentu merepresentasikan kemampuan generalisasi model terhadap data baru yang lebih kompleks, bervariasi, atau bebas dari bias kata kunci eksplisit. Secara keseluruhan, penelitian ini membuktikan bahwa *Logistic Regression* + BoW merupakan pendekatan yang efektif, efisien, dan interpretable untuk tugas klasifikasi topik berita, khususnya dalam konteks berita politik berbahasa Indonesia. Untuk pengembangan selanjutnya, disarankan melakukan validasi dengan dataset yang lebih besar dan beragam, menerapkan cross-validation, serta membandingkan performa dengan teknik representasi teks lain (seperti TF-IDF atau *word embeddings*) dan algoritma alternatif guna memastikan robustitas dan generalisasi model.

DAFTAR PUSTAKA

[1] B. Imran, M. N. Karim, and N. I. Ningsih, "Klasifikasi Berita Hoax Terkait Pemilihan Umum Presiden Republik Indonesia Tahun 2024 Menggunakan Naïve Bayes Dan Svm," *Din. Rekayasa*, vol. 20, no. 1, pp. 1–9, 2024, doi: 10.20884/1.dinarek.2024.20.1.27.

[2] J. Indrawan, R. E. Barzah, and H. Simanihuruk, "Instagram sebagai media komunikasi politik bagi generasi milenial," *Ekspresi dan Persepsi J. Ilmu Komun.*, vol. 6, no. 1, pp. 170–179, 2023.

[3] A. R. Hanum *et al.*, "Analisis Kinerja Algoritma Klasifikasi Teks Bert dalam Mendeteksi Berita Hoaks," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 3, pp. 537–546, 2024, doi: 10.25126/jtiik.938093.

[4] I Putu Gede Hendra Suputra, Linawati, I. G. Sukadarmika, and N. P. Sastra, "Klasifikasi Judul Berita Bahasa Indonesia Menggunakan Support Vector Machine Dan Seleksi Fitur Mutual Information," *J. Pendidik.*

Teknol. dan Kejuru., vol. 22, no. 1, pp. 69–79, 2025, doi: 10.23887/jptkundiksha.v22i1.89158.

[5] M. Zamzam, M. A. Kurniawan, and K. Khoiri, "Habaib Di Pusaran Kekuasaan: Studi Tentang Dinamika Politik Dan Agama Di Indonesia," *al-Akmal J. Stud. Islam*, vol. 3, no. 5, pp. 9–20, 2024.

[6] D. B. W. Alfredo Gormantara, "KLASIFIKASI KATEGORI DAN PELABELAN BERITA BAHASA INDONESIA MENGGUNAKAN MUTUAL INFORMATION DAN K- NEAREST NEIGHBORS," *J. Temat.*, vol. 8, pp. 75–82, 2020.

[7] R. Permana and F. A. Herdiana, "Analisis Klasifikasi Dan Prediksi Pola Publikasi Berita Pemprov DKI Jakarta Menggunakan Machine Learning," *J. Infortech*, vol. 7, no. 1, 2025.

[8] L. F. Chasanah and E. W. Pamungkas, "Klasifikasi Subjektif Berita Menggunakan Algoritma Machine Learning," in *Proceeding of Informatics Collaborations and Dessimenation Meeting*, 2025, pp. 140–143.

[9] Normah, B. Rifai, S. Vambudi, and R. Maulana, "Analisa Sentimen Perkembangan Vtuber Dengan Metode Support Vector Machine Berbasis SMOTE," *J. Tek. Komput. AMIK BSI*, vol. 8, no. 2, pp. 174–180, 2022, doi: 10.31294/jtk.v4i2.

[10] M. Fahmuddin, M. K. Aidid, and M. J. Taslim, "Implementasi Analisis Regresi Logistik Dengan Metode Machine Learning Untuk Mengklasifikasi Berita Di Indonesia," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 5, no. 03, pp. 155–162, 2023, doi: 10.35580/variansiunm116.

[11] Guruh Wijaya, Dudi Irawan, Zainul Arifin, Hardian Oktavianto, Miftahur Rahman, and Ginanjar Abdurrahman, "Studi Klasifikasi Topik Berita Dengan Algoritma Machine Learning," *J-Ensitec*, vol. 11, no. 01, pp. 10202–10206, 2024, doi: 10.31949/jensitec.v11i01.12037.

[12] N. E. Juliana, F. D. Khansa, A. M. H. Azis, R. I. Gunawan, and N. D. Cahya, "Klasifikasi Kategori Berita menggunakan Algoritma Support Vector Machine," *Gunung Djati Conf. Ser.*, vol. 3, 2021.

[13] I. F. Ramadhy and Y. Sibaroni, "Analisis Trending Topik Twitter dengan Fitur Ekspansi FastText Menggunakan Metode Logistic Regression," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 1, p. 1, 2022, doi: 10.30865/jurikom.v9i1.3791.

[14] A. Ananta Firdaus, A. Id Hadiana, and A. Kania Ningsih, "Klasifikasi Sentimen pada Aplikasi Shopee Menggunakan Fitur Bag of Word dan Algoritma Random Forest," *Ranah Res. J. Multidiscip. Res. Dev.*, vol. 6, no. 5, pp. 1678–1683, 2024, doi: 10.38035/rrj.v6i5.994.

[15] F. A. Wicaksono, A. Romadhony, and Hasmawati, "Sentiment Analysis of University Social Media Using Support Vector Machine and Logistic Regression Methods," *Ind. J. Comput.*, vol. 7, no. 2, pp. 15–24, 2022, doi: 10.34818/indojc.2022.7.2.638.