5th Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)
7 September 2024 – Jakarta, Indonesia
Volume 3, Nomor 2, September 2024 - ISSN 2962-8628 (*online*)

IMPLEMENTASI TEXT MINING UNTUK MENDETEKSI *HOAX* DENGAN MENGGUNAKAN *MULTINOMIAL NAÏVE BAYES* PADA STUDI KASUS PEMILU 2024

Dzulfikar Saif Assalam[1], Muhammad. Syafrullah[2*]

[1,2] Teknik Informatika, Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia

Email: [1]2011500770@student.budiluhur.ac.id, [2*]mohammad.syafrullah@budiluhur.ac.id
(* : corresponding author)

Abstrak-Pemilihan umum merupakan aspek penting dalam praktik demokrasi untuk menentukan pemimpin negara. Proses ini sangat dipengaruhi oleh kemajuan teknologi informasi, yang memudahkan penyebaran berita melalui media sosial dan platform *online*. Namun, kemudahan ini juga membuka peluang untuk penyebaran berita palsu atau *hoax*, yang dapat merugikan calon pemimpin dengan menyebarkan informasi yang salah dengan berbagai motif, seperti menakut-nakuti atau merusak reputasi. Permasalahan ini menjadi semakin kompleks karena berita hoax dapat mempengaruhi persepsi publik dan mengganggu proses demokratis, sering kali dimanfaatkan oleh pihak-pihak tertentu untuk mendapatkan keuntungan atau merusak citra seseorang. Untuk mengatasi permasalahan ini, penelitian ini mengembangkan sebuah sistem pendeteksi *hoax* menggunakan algoritma *Multinomial Naïve Bayes*. Algoritma ini merupakan metode klasifikasi yang sederhana namun efektif, mengklasifikasi teks berdasarkan frekuensi kemunculan kata untuk mengidentifikasi kebenaran informasi. Proses klasifikasi dalam sistem deteksi berita hoax ini melibatkan tahapan penting seperti preprocessing dan pembobotan kata menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk meningkatkan akurasi sistem dalam mengenali dan memfilter konten hoax. Sistem ini telah diuji coba dengan rasio pembagian data latih dan uji sebesar 80:20, menghasilkan akurasi deteksi 88%, presisi 93%, dan *recall* 86%. Hasil ini menunjukkan bahwa algoritma *Multinomial Naïve Bayes efektif* dalam mendeteksi berita *hoax* dan memiliki potensi besar dalam mendukung integritas informasi di era digital.

Kata kunci : *Text mining*, *hoax*, tf-idf, *multinomial naïve bayes*, pemilu 2024

IMPLEMENTATION OF TEXT MINING TO DETECT HOAXES USING MULTINOMIAL NAÏVE BAYES IN THE CASE STUDY OF THE 2024 ELECTIONS

Abstract-*Elections are a crucial aspect of democratic practices for determining national leaders. This process is greatly influenced by advancements in information technology, which facilitate the spread of news through social media and online platforms. However, this convenience also opens up opportunities for the dissemination of false or hoax news, which can harm political candidates by spreading misleading information for various motives, such as instilling fear or damaging reputations. This issue becomes increasingly complex as hoax news can affect public perception and disrupt the democratic process, often exploited by certain parties to gain advantage or tarnish someone's image. To address this issue, this study developed a hoax detection system using the Multinomial Naïve Bayes algorithm. This algorithm is a simple yet effective classification method, categorizing text based on the frequency of word occurrence to determine the veracity of information. The classification process in this hoax detection system involves critical stages such as preprocessing and word weighting using the Term Frequency-Inverse Document Frequency (TF-IDF) technique to enhance the system's accuracy in recognizing and filtering hoax content. The system was tested with a training and testing data ratio of 80:20, yielding a detection accuracy of 88%, precision of 93%, and recall of 86%. These results indicate that the Multinomial Naïve Bayes algorithm is effective in detecting hoax news and has significant potential to support information integrity in the digital era.*

Keywords*: Text mining, hoax, tf-idf, multinomial naïve bayes, pemilu 2024*

1. PENDAHULUAN

Pemilihan umum merupakan proses penting dalam negara demokratis untuk memilih wakil rakyat dan pemimpin negara. Indonesia, sebagai negara demokratis dengan populasi yang besar, menunjukkan keterlibatan aktif dalam proses demokrasi ini [1]. Menurut UUD 1945, pemilu di Indonesia dilaksanakan setiap lima tahun sekali dengan mengikuti prinsip-prinsip demokrasi yang telah ditetapkan.

Kemajuan dalam teknologi informasi telah mempermudah penyebaran informasi di era digital melalui media sosial dan berita online. Ada banyak informasi yang tersebar di internet, termasuk beberapa yang tidak benar dan berpotensi merugikan salah satu calon pemimpin negara. Media sosial memberikan kemampuan kepada

pengguna untuk berinteraksi, berbagi informasi, dan melakukan kegiatan lainnya melalui internet atau aplikasi, yang meningkatkan dinamika dalam komunikasi. Contoh platform media sosial termasuk YouTube, Facebook, blog, dan Twitter, yang memungkinkan pengguna untuk menyebarkan informasi positif maupun negatif, termasuk berita palsu atau *hoax* [2]. Informasi yang disebarkan di media sosial dan media massa saling terhubung secara dua arah, dan kemudahan dalam menyebarkan informasi ini kerap kali digunakan untuk menyebarluaskan berita yang menyesatkan atau palsu dengan tujuan tertentu.

Berita *hoax* adalah informasi yang disebarkan secara sengaja dan tidak benar dengan tujuan agar dipercaya sebagai kebenaran, berisi ujaran kebencian atau propaganda mencemarkan nama individu atau kelompok tertentu [3]. Berita *hoax* di Indonesia menyebar melalui berbagai media, termasuk media massa seperti cetak, elektronik, dan online. Di Indonesia, media sosial mendominasi sebagai platform utama penyebaran berita palsu dengan persentase sebesar 87.5%, signifikan lebih tinggi dibandingkan dengan aplikasi chatting (67%), situs web (28.2%), televisi atau radio (8.7%), media cetak (6.4%), dan email (2.6%) [2]. Berita *hoax* dapat merusak tatanan demokrasi serta stabilitas dalam kehidupan sosial, budaya, politik, dan ekonomi. Pemerintah, melalui Kepolisian Republik Indonesia, telah mengancam akan mengambil tindakan hukum terhadap mereka yang menyebarkan informasi palsu. Namun, tindakan ancaman tersebut dianggap sebagai pembatasan terhadap kebebasan berpendapat [4]. Sebagai upaya menghadapi berita *hoax*, Masyarakat Anti Fitnah Indonesia (MAFINDO) telah mendirikan situs turnbackhoax.id dan CekFakta.com.

Penelitian terdahulu yang dilakukan oleh Arizal Sabila Nurhikam, telah menggunakan algoritma seperti *Random Forest* dan *K-Nearest Neighbor* untuk mendeteksi berita palsu. Studi yang menguji algoritma Random Forest menggunakan data latih dan data uji dengan distribusi kelas yang seimbang, mencapai akurasi sebesar 84.88%, namun belum mengintegrasikan teknik pembobotan kata seperti TF-IDF yang dapat meningkatkan akurasi [5]. Sementara itu, penelitian yang dilakukan oleh I Wayan Santiyasa, menggunakan metode *K-Nearest Neighbor* menunjukkan pengaruh signifikan variasi nilai k terhadap akurasi, tetapi hanya mencapai akurasi maksimal 75.4% dan tidak membandingkan hasilnya dengan algoritma lain [6].

Penelitian sebelumnya telah berfokus pada penggunaan algoritma seperti *Random Forest* dan *K-Nearest Neighbor* untuk mendeteksi berita palsu, masih terdapat beberapa kesenjangan penelitian. Penelitian yang menggunakan algoritma *Random Forest* mencapai akurasi sebesar 84.88% dalam mendeteksi berita palsu terkait Pemilu 2024, Namun, penelitian tersebut belum menggunakan pembobotan kata seperti TF-IDF yang telah terbukti meningkatkan akurasi pada penelitian deteksi berita *hoax*. Sebaliknya, penelitian yang menggunakan algoritma *K-Nearest Neighbor* menunjukkan bahwa variasi nilai k berpengaruh signifikan terhadap akurasi deteksi hoax, tetapi penelitian tersebut hanya mencapai akurasi maksimal 75.4% dan juga tidak membandingkan hasilnya dengan algoritma lain yang mungkin lebih efektif.

Penelitian ini bertujuan untuk mengembangkan metode deteksi berita *hoax* yang lebih efektif dengan memanfaatkan algoritma *Multinomial Naïve Bayes* dan fitur pembobotan TF-IDF. Menggunakan data dari dataset MAFINDO yang dikumpulkan dari Maret 2023 hingga Mei 2024, dengan total 505 data, yang semuanya terkait dengan "pemilu 2024" sebagai kata kunci utama. Data yang diperoleh kemudian dilabeli secara otomatis berdasarkan kolom ‘status’, memudahkan analisis awal. Untuk mendeteksi berita *hoax* terkait Pemilu 2024, algoritma *Multinomial Naïve Bayes*, dengan pembobotan kata TF-IDF yang telah terbukti meningkatkan akurasi deteksi.

2. METODE PENELITIAN

Penelitian ini menggunakan *Multinomial Naïve Bayes* yang dikombinasikan dengan pembobotan TF-IDF untuk mendeteksi keberadaan berita *hoax*. Algoritma ini cukup efektif dalam mengklasifikasikan teks berdasarkan frekuensi kemunculan kata serta kemampuannya dalam mengatasi distribusi kata yang tidak merata dalam dokumen. Proses klasifikasi meliputi *preprocessing* data, pembagian dataset menjadi data latih dan data uji, pelatihan model, dan evaluasi model menggunakan akurasi, presisi, dan *recall*.

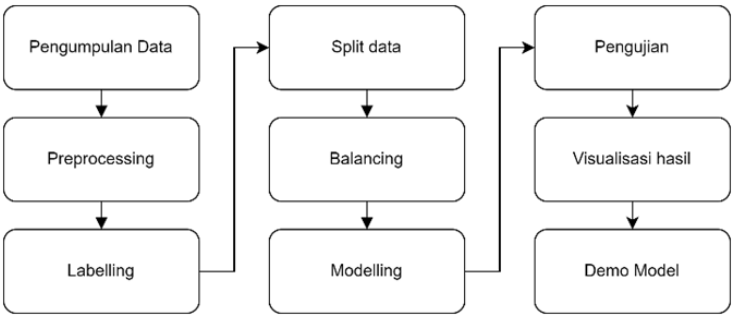
2.1 Data Penelitian

Penelitian ini menggunakan data dari dataset MAFINDO yang dikumpulkan pada Maret 2023 hingga Mei 2024. Proses pengumpulan data dilakukan menggunakan Python melalui teknik *API Data Fetching* dengan memanfaatkan *API Key* yang diberikan oleh MAFINDO. Dataset yang diperoleh didasarkan pada penggunaan kata kunci “pemilu 2024”. Jumlah total data yang terkumpul mencapai 505 data.

2.2 Penerapan Metode

Dalam penelitian ini, para peneliti mengembangkan sistem pendeteksi *hoax* menggunakan algoritma *Multinomial Naïve Bayes*, yang dirancang melalui serangkaian tahapan metodis yang terstruktur. Proses ini

dimulai dengan pengumpulan dan praproses data, diikuti oleh pelatihan algoritma dengan data yang telah disiapkan. Setelah sistem terlatih, dilakukan fase pengujian untuk mengukur efektivitasnya dalam mengidentifikasi berita *hoax*. Rincian lebih lengkap tentang setiap tahapan dalam proses pengembangan sistem ini dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Metode

Data diperoleh dengan menggunakan teknik *API Data Fetching* untuk mendapatkan dataset berupa berita yang sudah dilabeli dengan *hoax* dan *non-hoax*. Dataset diunggah ke dalam database untuk dilakukan *preprocessing*. Pada tahap preprocessing, dilakukan proses untuk membersihkan data dengan menyaring, memperbaiki dan menghapus kata-kata tertentu. Dalam penelitian ini, tahapan *preprocessing* yang dilakukan meliputi *case folding, cleansing, stopword removal, replacing slangword,* dan *stemming.*

a) *Case Folding*
Case Folding merupakan teknik yang mengonversi seluruh huruf dalam sebuah dokumen ke bentuk huruf kecil (*lower case*) untuk mengurangi duplikasi dalam data [7].

b) *Cleansing*
Cleansing adalah proses menghapus semua karakter non-alfabet dalam dokumen untuk menghilangkan karakter yang tidak diinginkan dan tidak bermakna. Karakter tersebut mencakup angka, tanda #, @, emoji, dan tautan dari situs web yang terdapat dalam dokumen [8].

c) *Stopword Removal*
Stopword Removal adalah metode eliminasi kata-kata umum dari teks yang digunakan dalam analisis bahasa alami atau pemrosesan teks. Tujuan dari proses ini adalah untuk membuang kata-kata yang dianggap tidak penting, termasuk preposisi, kata benda, dan kata-kata yang sering muncul lainnya [9]

d) *Replacing Slangword*
Slangword adalah istilah yang merujuk pada kata-kata yang tidak mematuhi aturan ejaan resmi bahasa Indonesia (EYD). Kategori ini mencakup singkatan, kata-kata gaul atau modern, serta kesalahan penulisan [10]

e) *Stemming*
Stemming adalah metode penghilangan prefiks (imbuhan awal) dan sufiks (imbuhan akhir) pada kata. Prefiks di kamus bahasa Indonesia meliputi me-, mem-, meng-, di-, dan lain-lain. Sufiks mencakup -i, -nya, -an. Hasil dari proses ini adalah kata dasar tanpa imbuhan [11].

Setelah dilakukan *preprocessing*, dilakukan tahap *labelling*, di mana dokumen atau kalimat dikategorikan berdasarkan ciri atau karakteristik yang ada di dalamnya. Proses pelabelan data ini dilakukan secara langsung dari dataset yang terdapat di kolom 'status'. Setelah itu, dilakukan *split data*. Dalam proses *split data*, dataset dibagi menjadi dua bagian: data latih dan uji dengan beberapa rasio, yaitu 90:10, 80:20, 70:30, 60:40 dan 50:50, *default*nya adalah 80:20, dengan 80% sebagai data latih dan 20% sebagai data uji untuk memastikan model dilatih secara efektif. Selanjutnya, dalam proses *balancing*, data latih diimbangi antara data *hoax* dan *non-hoax* menggunakan teknik *random undersampling* untuk mencegah *overfitting*.

Proses *modelling* melibatkan seleksi data latih, tokenisasi, pengumpulan kata, perhitungan vektor token, *term frequency, inverse document frequency*. *Term Frequency* adalah metode yang menentukan bobot kata dalam dokumen berdasarkan frekuensi kemunculannya. *Inverse Document Frequency* (IDF) adalah untuk menurunkan bobot kata yang sering muncul di banyak dokumen [12].

Rumus perhitungan TF-IDF:

$$TF(d,t) = \frac{f_{t,d}}{\sum t,d} \quad (1)$$

$$IDF(t) = \log\left(\frac{N_d}{df(t)}\right) \quad (2)$$

$$TF - IDF = TF(d,t) \times IDF(t) \quad (3)$$

Tahap selanjutnya adalah *Multinomial Naïve Bayes*. Multinomial Naïve Bayes adalah pengembangan dari algoritma Bayes yang cocok untuk klasifikasi teks atau dokumen [2]. Dalam rumus Multinomial Naïve Bayes, kelas dokumen ditentukan tidak hanya oleh kehadiran kata, tetapi juga oleh frekuensi kemunculannya [12]. Proses klasifikasi dilakukan setelah model pelatihan berhasil dibuat.

Rumus *Multinomial Naïve Bayes:*

$$P(c,d) = \frac{N_c}{N} \times P(t_1,c) \times \ldots \times P(t_n,c) \quad (4)$$

Rumus untuk probabilitas kata ke-n yang digunakan dengan pembobotan kata TF-IDF adalah sebagai berikut:

$$P(t_n,c) = \frac{Wct + 1}{(\sum W, €VW, ct) + B'} \quad (5)$$

Proses klasifikasi ini dibagi menjadi tiga langkah. Pertama, model yang telah disimpan dalam format JSON dimuat ke dalam sistem. Model ini mengandung probabilitas kelas prior dan likelihood yang diperoleh dari data pelatihan menggunakan metode TF-IDF. Kedua, sistem melakukan perbandingan probabilitas total untuk setiap kelas berdasarkan data yang sedang diuji, mengintegrasi probabilitas likelihood dan prior dari kata-kata dalam dokumen tersebut. Ketiga, probabilitas untuk setiap token dihitung untuk menentukan klasifikasi akhir dokumen, dengan memilih kelas yang memiliki probabilitas tertinggi. Proses ini memungkinkan penentuan '*hoax'* atau '*non-hoax*'.

Pada peneltian ini, pengujian dilakukan dengan *confusion matrix*. *Confusion Matrix* adalah langkah dalam analisis dan evaluasi kinerja sistem yang telah dirancang. Kinerja ini diukur menggunakan nilai akurasi, presisi, dan recall. *Confusion Matrix* digunakan sebagai metode untuk menghitung nilai akurasi [2].

Tabel 1. *Confusion Matrix*

Aktual	Prediksi	
	Hoax	*Non-Hoax*
Hoax	*TP (True Positive)*	FN (*False Negative)*
Non-Hoax	FP (*False Positive)*	TN (*True Negative)*

Pada tahap ini, performa model yang telah dibangun akan dievaluasi dengan menghitung akurasi, presisi, dan recall. Akurasi merupakan ukuran dari seberapa sesuai nilai prediksi dengan nilai aktual. Presisi mengukur seberapa tepat jawaban yang diberikan oleh sistem sesuai dengan yang diminta oleh pengguna. Sedangkan *recall* menilai sejauh mana sistem berhasil menemukan kembali informasi yang relevan.

$$Akurasi = \frac{TP + TN}{TP + FP + TN + FN} \quad (6)$$

$$Presisi = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Dataset penelitian ini adalah data MAFINDO. Proses pengambilan data yang dilakukan melalui *teknik API Data Fetching* menggunakan bahasa pemrograman python. Data diambil pada tanggal 20 Mei 2024, diperoleh sebanyak 505 data pada rentang waktu Maret 2023 – Mei 2024. Dari hasil pengumpulan data terdapat 272 *hoax* dan 233 *non-hoax*. Kata kunci yang digunakan adalah "pemilu 2024".

Tabel 2. *Sample Dataset*

No	*Author*	*Title*	Status
1	Kompas	[HOAKS] Prabowo-Gibran Batal Dilantik oleh MPR	*Hoax*
2	Mafindo	PENEMUAN GUDANG KOTAK SUARA GANDA DI MAKASSAR UNTUK KECURANGAN PEMILU 2024	*Hoax*
3	Tempo	Menyesatkan, Klaim bahwa Jokowi Tak Terlibat atau Tak Ikut Kampanye Kandidat Manapun saat Pilpres 2024	*Hoax*
4	Tempo	Benar, Video Seorang Pria Memasukkan Surat Suara dari Kantong Plastik ke Kotak Suara di TPS Pidie, Aceh	*Non-Hoax*
5	Tempo	Benar, Video tentang WNI di Taipei Lapor Surat Suara Dalam Amplop Telah Tercoblos	*Non-Hoax*

3.2 *Preprocessing* dan *Labelling*

Pada tahap ini, data mentah akan melalui proses penghapusan data yang tidak relevan untuk memastikan kualitas data yang lebih baik. Data yang dianggap tidak penting atau tidak mendukung analisis lebih lanjut akan dihilangkan. Selain itu, data yang ada juga akan dimodifikasi ke dalam format yang lebih sederhana dan konsisten, sehingga memudahkan dalam proses analisis berikutnya.

Tabel 3. Tahapan *Preprocessing*

Tahapan *Preprocessing*	Hasil
Teks Asli	PENEMUAN GUDANG KOTAK SUARA GANDA DI MAKASSAR UNTUK KECURANGAN PEMILU 2024
Case Folding	penemuan gudang kotak suara ganda di makassar untuk kecurangan pemilu 2024
Cleansing	penemuan gudang kotak suara ganda di makassar untuk kecurangan pemilu
Stopword Removal	penemu gudang kotak suara gand makassar kecurangan pemilu
Replacing Slangword	penemu gudang kotak suara gand makassar curang pemilu
Stemming	temu gudang kotak suara ganda makassar curang pemilu

Proses pelabelan dilakukan secara langsung dari dataset yang telah tersedia, di mana setiap entri sudah memiliki kolom ‘status’ yang berisi label ‘*hoax’* atau ‘*non-hoax’*. Label ini memberikan informasi penting untuk membedakan antara berita palsu dan berita yang valid. Dengan adanya label ini, proses analisis dan klasifikasi data menjadi lebih terarah dan efisien.

3.3 *Split data* dan *Balancing*

Setelah data dilakukan *preprocessing*, pembagian data dilakukan dengan beberapa rasio yaitu 90:10, 80:20, 70:30, 60:40, dan 50:50, dengan rasio *default* adalah 80:20. Dari 505 data yang tersedia, 80% digunakan sebagai data latih dan 20% sebagai data uji. Dengan demikian, 404 data digunakan untuk latih dan 101 data untuk uji. Setelah pembagian, data tersebut disimpan kembali ke dalam database untuk proses selanjutnya.

Tabel 4. Tabel *Split Data* Latih dan Uji

	Rasio	**Data Latih**	**Data Uji**
505 Data	Rasio 90:10	454	51
	Rasio 80:20	404	101
	Rasio 70:30	353	152
	Rasio 60:40	303	202
	Rasio 50:50	252	253

Setelah pembagian data, dilakukan proses *balancing*. Dari 404 data latih, terdapat 211 data *hoax* dan 193 data *non-hoax*, yang menunjukkan ketidakseimbangan. Oleh karena itu, proses *balancing* diperlukan agar pemodelan lebih efektif. Setelah proses *balancing*, data latih menjadi seimbang dengan masing-masing kategori memiliki 193 data *hoax* dan 193 data *non-hoax*.

3.4 Pembobotan TF-IDF

Setelah data melalui tahap *split data* dan *balancing*. langkah berikutnya adalah pembobotan kata menggunakan metode TF-IDF. Tahap ini melibatkan perhitungan *Term Frequency* (TF), *Inverse Document Frequency* (IDF), dan mengombinasikan kedua nilai tersebut untuk memperoleh nilai TF-IDF. Pembobotan dengan TF-IDF dilakukan menggunakan data latih.

Tabel 5. Sampel Data Latih

Dokumen	Teks	Label
D1	prabowo capai persen suara hasil pemilu taiwan	*Hoax*
D2	yusril sebut prabowo gibran diskualifikasi	*Hoax*
D3	kades ntb amuk massa laku curang pemilu	*Hoax*
D4	masa tenang pemilu apk jalan kota bogor tertib	*Non-Hoax*
D5	surat suara pemilu coblos ilegal malaysia	*Non-Hoax*
D6	bawaslu sebut tps indonesia sulit jangkau	*Non-Hoax*

Selanjutnya, dilakukan perhitungan TF-IDF (*Term Frequency-Inverse Document Frequency*) untuk mengukur seberapa penting sebuah kata dalam dokumen tertentu relatif terhadap keseluruhan korpus. TF-IDF membantu mengidentifikasi kata-kata yang memiliki bobot informasi tinggi dan relevan untuk analisis, dengan menghitung frekuensi kemunculan kata dalam dokumen (TF) dan mengurangi bobot kata-kata umum yang sering muncul di banyak dokumen (IDF). Hasil perhitungan TF-IDF ini kemudian digunakan sebagai fitur dalam model pembelajaran mesin untuk meningkatkan akurasi prediksi.

Tabel 6. Perhitungan TF-IDF

	TF-IDF D1	TF-IDF D2	TF-IDF D3	TF-IDF D4	TF-IDF D5	TF-IDF D6
prabowo	0.068	0.095	0	0	0	0
capai	0.111	0	0	0	0	0
persen	0.111	0	0	0	0	0
suara	0.068	0	0	0	0.080	0
hasil	0.111	0	0	0	0	0
pemilu	0.025	0	0.025	0.022	0.029	0
taiwan	0.111	0	0	0	0	0
yusril	0	0.156	0	0	0	0
gibran	0	0.095	0	0	0	0.080
diskualifikasi	0	0.156	0	0	0	0
kades	0	0.156	0	0	0	0
ntb	0	0	0.111	0	0	0
amuk	0	0	0.111	0	0	0
massa	0	0	0.111	0	0	0
laku	0	0	0.111	0	0	0
curang	0	0	0.111	0	0	0
masa	0	0	0.111	0	0	0
tenang	0	0	0	0.097	0	0
apk	0	0	0	0.097	0	0
jalan	0	0	0	0.097	0	0
kota	0	0	0	0.097	0	0
bogor	0	0	0	0.097	0	0
tertib	0	0	0	0.097	0	0
surat	0	0	0	0.097	0	0
coblos	0	0	0	0	0.130	0
ilegal	0	0	0	0	0.130	0
malaysia	0	0	0	0	0.130	0
bawaslu	0	0	0	0	0.130	0
tps	0	0	0	0	0	0.130

indonesia	0	0	0	0	0	0.130
sulit	0	0	0	0	0	0.130
jangkau	0	0	0	0	0	0.130

Selanjutnya, nilai TF-IDF dari setiap kata kemudian dijumlahkan sesuai dengan masing-masing label, yaitu '*hoax*' dan '*non-hoax*'. Dengan menjumlahkan nilai TF-IDF ini, kita dapat mengidentifikasi kata-kata mana yang lebih sering muncul dan memiliki bobot lebih besar dalam kategori tertentu.

Tabel 7. Perhitungan TF-IDF Per Kelas

	Hoax	*Non-Hoax*
prabowo	0.164	0
capai	0.111	0
persen	0.111	0
suara	0.068	0.080
hasil	0.111	0
pemilu	0.050	0.051
taiwan	0.111	0
yusril	0.156	0
gibran	0.095	0.080
diskualifikasi	0.156	0
kades	0.156	0
ntb	0.111	0
amuk	0.111	0
massa	0.111	0
laku	0.111	0
curang	0.111	0
masa	0.111	0
tenang	0	0.097
apk	0	0.097
jalan	0	0.097
kota	0	0.097
bogor	0	0.097
tertib	0	0.097
surat	0	0.097
coblos	0	0.130
ilegal	0	0.130
malaysia	0	0.130
bawaslu	0	0.130
tps	0	0.130
indonesia	0	0.130
sulit	0	0.130
jangkau	0	0.130
Total	**1.956**	**2.059**

3.5 *Multinomial Naïve Bayes*

Setelah memperoleh hasil pembobotan TF-IDF, langkah berikutnya adalah menghitung probabilitas kelas prior, yang digunakan untuk menentukan probabilitas awal dari setiap kelas dalam dataset. Probabilitas kelas prior ini dihitung dengan membagi jumlah dokumen dalam setiap kelas dengan total dokumen yang tersedia. Dari data latih yang telah disediakan, terdapat enam contoh data, di mana tiga di antaranya berlabel *hoax* dan tiga lainnya berlabel *non-hoax*. Oleh karena itu, probabilitas kelas prior untuk setiap kelas adalah sebagai berikut:

$$Probabilitas\ Kelas\ Prior\ Hoax\ = \frac{Jumlah\ Dokumen\ Hoax}{Total\ Dokumen} \tag{9}$$

$$Probabilitas\ Kelas\ Prior\ Hoax\ = \frac{3}{6} = 0.5$$

$$Probabilitas\ Kelas\ Prior\ Non-Hoax\ = \frac{Jumlah\ Dokumen\ Non-Hoax}{Total\ Dokumen} \tag{10}$$

$$Probabilitas\ Kelas\ Prior\ Non-Hoax = \frac{3}{6} = 0.5$$

Setelah didapatkan probabilitas prior untuk setiap kelas, langkah berikutnya adalah melakukan proses pengujian dengan menggunakan satu sampel data uji. Dalam proses ini, sampel data uji akan dianalisis berdasarkan nilai TF-IDF yang sudah dihitung sebelumnya. Probabilitas dari setiap kelas (*hoax* dan *non-hoax*) kemudian dihitung dengan menggabungkan probabilitas prior dan kontribusi dari fitur-fitur (kata-kata) dalam sampel tersebut.

Tabel 8. Sampel Data Uji

Teks	Label
kades ntb diskualifikasi curang masa tenang	*Hoax*

Selanjutnya, hitung probabilitas *likelihood* dan posterior setiap kata dalam data uji untuk setiap kelas. Tabel 9 dan 10 menampilkan probabilitas setiap kata di kelas *hoax* dan kelas *no-hoax*.

Tabel 9. Probabilitas Kelas *Hoax*

Kata	**Probabilitas *Likelihood***	**Probabilitas Posterior**
kades	prob = (0+0+0.11+1) / (1.96+24.17) = **0.043**	
ntb	prob = (0+0+0.11+1) / (1.96+24.17) = **0.043**	
diskualifikasi	prob = (0+0+0.16+1) / (1.96+24.17) = **0.044**	0.5 × 0.043 × 0.043 × 0.044 × 0.043 × 0.038 × 0.038 = **0.000000002490587**
curang	prob = (0+0+0.11+1) / (1.96+24.17) = **0.043**	
masa	prob = (0+0+0+1) / (1.96+24.17) = **0.038**	
tenang	prob = (0+0+0+1) / (1.96+24.17) = **0.038**	

Tabel 10. Probabilitas Kelas *Non-Hoax*

Kata	**Probabilitas *Likelihood***	**Probabilitas Posterior**
kades	prob = (0+0+0.11+1) / (1.96+24.17) = **0.043**	
ntb	prob = (0+0+0.11+1) / (1.96+24.17) = **0.043**	
diskualifikasi	prob = (0+0+0.16+1) / (1.96+24.17) = **0.044**	0.5 × 0.038 × 0.038 × 0.038 × 0.038 × 0.042 × 0.042 = **0.000000001847451**
curang	prob = (0+0+0.11+1) / (1.96+24.17) = **0.043**	
masa	prob = (0+0+0+1) / (1.96+24.17) = **0.038**	
tenang	prob = (0+0+0+1) / (1.96+24.17) = **0.038**	

Berdasarkan hasil perhitungan, probabilitas data uji untuk kelas *hoax* adalah **0.000000002490587**, sedangkan untuk kelas *non-hoax* adalah **0.000000001847451**. Dari hasil perbandingan probabilitas tersebut, dokumen tersebut lebih cenderung dikategorikan ke dalam kelas *hoax*, karena probabilitasnya lebih besar daripada kelas *non-hoax*. Hal ini menunjukkan bahwa dokumen tersebut mengandung elemen atau kata-kata yang lebih sering terkait dengan informasi palsu.

3.6 Pengujia

Pengujian dilakukan untuk mengonfirmasi bahwa sistem yang diimplementasikan sesuai dengan spesifikasi yang sudah ditentukan sebelumnya. Pengujian bertujuan untuk mengukur akurasi, presisi, dan *recall* dalam implementasi algoritma *Multinomial Naïve Bayes* untuk memprediksi label pada data testing. Penelitian ini juga membandingkan hasil dari berbagai rasio perbandingan data, yaitu 90:10, 80:20, 70:30, 60:40, dan 50:50.

Tabel 10. Pengujian *Confusion Matrix*

Rasio	**TP**	**TN**	**FP**	**FN**
Rasio 90:10	26	17	3	4
Rasio 80:20	53	36	4	8
Rasio 70:30	76	47	13	15
Rasio 60:40	93	61	24	24
Rasio 50:50	107	72	38	35

Tabel 11. Hasil *Confusion Matrix* di Setiap Rasio

Pengujian			
Rasio	**Akurasi**	**Presisi**	***Recall***
Rasio 90:10	86%	89%	86%
Rasio 80:20	88%	93%	86%
Rasio 70:30	81%	85%	83%
Rasio 60:40	76%	79%	79%
Rasio 50:50	71%	73%	75%

Berdasarkan tabel pengujian di atas, terlihat bahwa hasil pengujian terbaik terdapat di rasio 80:20 dengan nilai akurasi 88%, presisi 93%, dan *recall* 86%, temuan menunjukkan bahwa algoritma *Multinomial Naïve Bayes* cukup baik, dengan performa yang menunjukkan keandalan dalam deteksi *hoax*.

4. KESIMPULAN

Berdasarkan analisis terhadap 505 data, terdapat 272 data *hoax* dan 233 data *non-hoax*, menunjukkan mayoritas masyarakat Indonesia masih rentan terhadap penyebaran informasi palsu menjelang Pemilu 2024. Hasil pembobotan TF-IDF dan penerapan algoritma *Multinomial Naïve Bayes* dengan rasio 80:20 menghasilkan performa terbaik, yaitu akurasi 88%, presisi 93%, dan *recall* 86%. Beberapa tahap utama dalam penelitian ini mencakup pengambilan data, *preprocessing*, *labelling*, *modelling*, *split data*, *balancing*, dan klasifikasi dengan *Multinomial Naïve Bayes*. Tahapan *preprocessing* yang dilakukan dengan baik berperan penting dalam menghasilkan hasil yang optimal, sementara penggunaan dataset MAFINDO membantu dalam memberikan label pada data serta mengurangi waktu dan usaha dalam proses *labelling*.

DAFTAR PUSTAKA

[1] Chely Aulia Misrun, E. Haerani, M. Fikry, and E. Budianita, "Analisis sentimen komentar youtube terhadap Anies Baswedan sebagai bakal calon presiden 2024 menggunakan metode naive bayes classifier," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 4, no. 1, pp. 207–215, 2023, doi: 10.37859/coscitech.v4i1.4790.

[2] C. S. Sriyano and E. B. Setiawan, "Pendeteksian Berita Hoax Menggunakan Naive Bayes Multinomial Pada Twitter dengan Fitur Pembobotan TF-IDF," *e-Proceeding of Engineering*, vol. 8, no. 2, 2021, pp. 3396-3405.

[3] V. O. Yamin *et al.*, "Penerapan Naïve Bayes Classifier dengan Algoritma Nazief dan Adriani Untuk Deteksi Hoaks," Prosiding Sempatin, vol. 1, no. 1, 2023, pp. 335-344.

[4] M. Ibrahim, E. Bu'ulolo, and I. Lubis, "Penerapan Algoritma Naive Bayes Classifier Untuk Mendeteksi Tingkat Krediblitas Hoax News/ Fake News Pada Sosial Media Di Indonesia Berbasis Android (Studi Kasus : Kantor Tribun Medan)," *RESOLUSI: Rekayasa Teknik Informatika dan Informasi*, vol. 1, no. 1, pp. 9-17, 2020, [Online]. Available: https://djournals.com/resolusi

[5] A. S. Nurhikam *et al.*, "Deteksi Berita Palsu Pada Pemilu 2024 Dengan Menggunakan Algoritma Random Forest," vol. 7, no. 1, pp. 41–50, 2023, [Online]. Available: http://e-journal.unipma.ac.id/index.php/doubleclick

[6] I. W. Santiyasa, et al, "Identification of Hoax Based on Text Mining Using K-Nearest Neighbor Method," *Jurnal Elektronik Ilmu Komputer Udayana*, vol. 10, no. 2, pp. 217-226, 2021.

[7] P. Sarah Kayaningtias and P. Pandu Adikara, "Analisis Sentimen Angket Kepuasan Pasien Puskesmas menggunakan Metode Improved K-Nearest Neighbor," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 6, no. 3, pp. 1138-1148, 2022.

[8] A. A. Aqsa, and L. Syafie, "Buletin Sistem Informasi dan Teknologi Islam Perbandingan Metode Naïve Bayes dan SVM dalam Analisis Sentimen Netizen Twitter Terhadap Isu Kemenkeu," *BUSITI: Buletin Sistem Informasi dan Teknologi Islam*, vol. 4, no. 4, pp. 327–338, 2023.

[9] B. Imran, M. Nasirudin Karim, and N. Isna Ningsih, "Klasifikasi Berita Hoax Terkait Pemilihan Umum Presiden Republik Indonesia Tahun 2024 Menggunakan Naïve Bayes dan Svm," *Dinamika Rekayasa*, vol. 20, no. 1, pp. 1-9, 2024.

[10] A. Ilham and W. Pramusinto, "Analisis Sentimen Masyarakat Terhadap Kesehatan Mental Pada Twitter Menggunakan Algoritma K-Nearest Neighbor," *Prosiding Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)*, vol. 2, no. 2, 2023, pp. 539-547.

[11] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," vol. 8, no. 1, pp. 147–156, 2021, doi: 10.25126/jtiik.202183944.

[12] F. A. Ramadhan, S. H. Sitorus, and T. Rismawan, "Penerapan Metode Multinomial Naïve Bayes untuk Klasifikasi Judul Berita Clickbait dengan Term Frequency - Inverse Document Frequency," *Jurnal Sistem dan Teknologi Informasi (JustIN)*, vol. 11, no. 1, p. 70, Jan. 2023, doi: 10.26418/justin.v11i1.57452.