JURNAL FASILKOM
Volume 14 No. 2 | Agustus 2024: 410-419

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Deteksi Berita Salah Pada Pemilihan Umum Presiden 2024 Menggunakan Metode *Naïve Bayes* Berbasis Website

Aziz Musthafa[1], Dihin Muriyatmoko[2], Taufiqurrahman[3], Surya Kamal Sholihin[4]
[1,2,3,4]Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Darussalam Gontor
[1]aziz@unida.gontor.ac.id, [2]dihin@unida.gontor.ac.id, [3]taufiqurrahman@unida.gontor.ac.id,
[4]suryakamalsholihin33@mhs.unida.gontor.ac.id*

Abstract

Approaching the general election period, a lot of false news is emerging among the public, leading public opinion to vote for presidential candidates who are supported by false news makers. At the beginning of 2024, the Ministry of Communication and Information has identified a total of 203 election hoax issues spread across various digital news platforms. Therefore, it makes people who want to follow news about election developments become doubtful. The aim of this research is to create a machine learning platform that can classify news as true or false automatically and easily. In classifying news, Text Mining techniques are used which can process text or document data to obtain the required information. The method used is Naïve Bayes classification. The data used is true news and false news from the Turn Back Hoax site by MAFINDO (Indonesian Anti-Defamation Society) which provides news sources that are verified as true and have labeled false news circulating in the community. Implementation using a website-based application for classification of General Election news. The classification results from the website using the Naïve Bayes classification method obtained good accuracy evaluation results, namely 91% classification accuracy level. Based on the test results, it is hoped that the results of this research can contribute to people's digital literacy regarding the accuracy of election news.

Keywords: *False News, Classification, Text Mining, Naïve Bayes, Machine Learning.*

Abstrak

Mendekati masa pemilihan umum, banyak berita salah yang muncul ditengah-tengah masyarakat menggiring opini masyarakat agar memilih calon presiden yang didukung pembuat berita salah. Diawal tahun 2024, Kementerian Komunikasi dan Informasi telah mengidentifikasi total 203 isu *hoax* pemilu yang tersebar di berbagai platform berita digital. Oleh karena itu membuat masyarakat yang ingin mengikuti berita perkembangan pemilu menjadi ragu. Tujuan dari penelitian ini membuat aplikasi pembelajaran mesin yang dapat mengklasifikasikan berita benar atau salah secara otomatis dan mudah. Dalam mengklasifikasikan berita, digunakan teknik Penambangan Teks (*Text Mining*) yang dapat mengolah data teks atau dokumen untuk mendapatkan informasi yang dibutuhkan. Metode yang digunakan yaitu klasifikasi *Naïve Bayes*. Data yang digunakan berupa berita benar dan berita salah dari situs *Turn Back Hoax* oleh MAFINDO (Masyarakat Anti Fitnah Indonesia) yang menyediakan sumber berita terverifikasi benar dan telah melabeli berita salah yang beredar di masyarakat. Implementasi menggunakan aplikasi berbasis website untuk klasifikasi berita Pemilihan Umum. Hasil klasifikasi dari website dengan menggunakan metode klasifikasi *Naïve Bayes* mendapatkan hasil evaluasi akurasi yang baik, yaitu sebesar 91% tingkat akurasi klasifikasinya. Berdasarkan hasil pengujian tersebut, diharapkan hasil penelitian ini dapat memberikan sumbangan bagi literasi digital masyarakat mengenai kearuratan berita pemilu.

Kata Kunci: Berita salah, Klasifikasi, Penambangan Teks, *Naïve Bayes, Machine Learning.*

1. Pendahuluan

Saat ini, teknologi terus berkembang pesat untuk memudahkan segala kebutuhan setiap orang. Perkembangan teknologi modern ini juga didukung oleh kemajuan teknologi informasi ke era digital, yang memudahkan pengguna internet untuk mendapatkan informasi terbaru dari seluruh dunia. Berita-berita dari berbagai belahan dunia kini sudah masuk ke tahap digital dan dapat diakses melalui website berita online seperti Kompas, Tribun News, Detik.com, CNN, dan lain sebagainya.

Kemudahan menyebarkan informasi tidak hanya berhenti pada digitalisasi berita saja. Pada tahun 2004, muncul istilah Web 2.0 yang menjelaskan cara baru di mana developer dan pengguna memanfaatkan website sebagai platform yang kontennya tidak lagi dibuat dan dipublikasikan secara individual oleh developer saja, tetapi dapat dimodifikasi secara fleksibel oleh seluruh

P-ISSN : 2089-3353
E-ISSN : 2808-9162

pengguna. Web 2.0 mencakup media sosial yang sekarang menjadi kebutuhan wajib bagi setiap orang, baik untuk keperluan pribadi maupun bisnis. Munculnya platform Web 2.0 seperti Blogspot, Facebook, Twitter (X), YouTube, dan lain sebagainya memudahkan setiap orang untuk berbagi informasi pribadi atau kelompok kepada publik [1].

Meskipun penyebaran informasi menjadi lebih mudah, hal ini membuka celah bagi kelompok-kelompok jahat untuk menyebarkan berita palsu atau hoaks. Berita palsu adalah berita yang sengaja dibuat dan disebarkan untuk menyesatkan pembaca. Berita palsu memiliki dua ciri utama: pertama, mengandung informasi yang salah; kedua, dibuat dengan niat tidak jujur untuk menyesatkan pembaca [2].

Menjelang Pemilu, berbagai kelompok memanfaatkan berita palsu untuk kepentingan masing-masing. Kementerian Kominfo telah menemukan 101 isu berita palsu tentang Pemilu yang beredar sejak Januari 2023 hingga 26 Oktober 2023, meningkat hampir sepuluh kali lipat dibandingkan tahun sebelumnya yang hanya ada 10 berita palsu. Hingga awal 2024, Kominfo telah menangani total 203 isu hoaks Pemilu yang tersebar di berbagai platform digital. Peningkatan ini berpotensi merugikan kelangsungan Pemilu dan perlu diperhatikan dengan serius, karena keberadaan berita palsu mengenai Pemilu tidak hanya merusak kualitas demokrasi, tetapi juga dapat memecah belah bangsa [3].

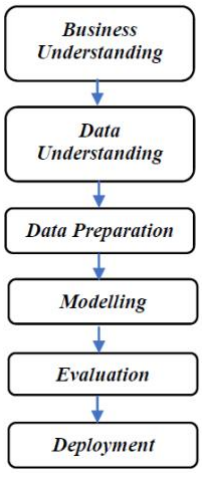
Penelitian ini bertujuan untuk membuat program yang dapat mengklasifikasikan berita benar atau palsu secara otomatis. Penelitian ini menggunakan metode Penambangan Teks (Text Mining) yang mengolah data teks atau dokumen untuk mendapatkan informasi yang dibutuhkan. Metode klasifikasi yang digunakan adalah Naïve Bayes, yang telah terbukti stabil dalam mengklasifikasikan data teks. Data yang digunakan berupa berita benar dan berita palsu, diambil dari situs Turn Back Hoax dari MAFINDO (Masyarakat Anti Fitnah Indonesia) yang menyediakan sumber berita benar dan telah melabeli berita palsu yang beredar di masyarakat. Diharapkan penelitian ini dapat membantu mengurangi penyebaran berita palsu pada masa Pemilu 2024.

Penelitian yang dilakukan oleh (Fani Prasetya dan Ferdiansyah) dengan judul "Analisis Data Mining Klasifikasi berita Hoax COVID 19 Menggunakan Algoritma *Naive Bayes*". Menguji algoritma klasifikasi *naïve bayes* untuk mengklasifikasikan berita salah COVID 19. Hasil penelitian yang telah dilakukan dengan menggunakan model klasifikasi naïve bayes dan evaluasi cross validation dapat melakukan klasifikasi berita salah dengan baik, akurasi yang dihasilkan sebesar 86.3%. Data yang diprediksi salah juga tidak terlalu banyak dari total 300 dataset hanya 41 yang dinyatakan salah dalam pelabelan tidak sampai 2% dari keseluruhan total dataset [4].

Penelitian yang dilakukan oleh (Muhammad Fadhil Muttaqin, dkk) membuat model klasifikasi berita palsu virus Covid-19 dengan menerapkan algoritma Support Vector Machine (SVM). Hasil penelitian menyatakan bahwa model klasifikasi berita palsu Covid-19 dengan Algoritma Support Vector Machine (SVM) memiliki nilai akurasi 78%. Setelah uji coba model SVM terhadap dataset Covid-19, deploy model dilakukan dan diharapkan untuk masa yang akan datang masyarakat dapat mengakses situs pengecekan berita palsu. Selain mendapatkan prediksi judul berita yang dikategorikan dengan "Fake" dan "True", web juga dapat menampilkan persentase probabilitas dari prediksi yang dilakukan oleh model. Dengan adanya sistem ini, dapat memprediksi judul berita tentang virus Covid-19 seakurat mungkin, baik berita palsu maupun berita yang berisikan fakta [5].

2. Metode Penelitian

Penelitian ini bertujuan membuat program berbasis web yang dapat mengklasifikasikan berita kedalam berita salah dan berita benar menggunakan Text Mining dengan metode klasifikasi Naïve Bayes. Pada penelitian sebelumnya mengenai "Implementasi *Modified Enhanced Confix Stripping Stemmer* pada Klasifikasi Fake News Covid-19" mendapat hasil akurasi yang bagus dalam mengklasifikasikan data teks. Metode penelitian klasifikasi berita PEMILU 2024 menggunakan tahapan penelitian CRIPS-DM, yang mana terdiri dari urutan penelitian pada Gambar 1 sebagai berikut:



Gambar 1 Metode Penelitian CRIPS-DM [6]

2.1 Bussiness Understanding

Adapun inti permasalahan dalam penelitian ini adalah berita salah yang tersebar di masyarakat selama masa PEMILU 2024 sesuai dengan hasil identifikasi Kementerian Kominfo. Sedangkan tujuan dari penelitian ini adalah mengklasifikasi berita yang tersebar apakah termasuk berita benar atau berita salah.

Dalam tahapan ini juga mengkaji literatur mengenai teori dalam *text mining*, metode klasifikasi *Naïve Bayes*, serta teknik klasifikasi berita benar dan berita salah yang diharapkan dapat memberikan hasil yang optimal.

2.2 Data Understanding

Proses berikutnya adalah mempelajari data dengan mengumpulkan data yang dibutuhkan terlebih dahulu. Untuk pengumpulan data, penelitian ini menggunakan teknik web *Scraping* dengan python menggunakan *BeautifulSoup*, yaitu teknik pengambilan data dari sebuah halaman website.

Proses *scraping* terdiri dari beberapa langkah menggunakan perpustakaan *BeautifulSoup,* langkah pertama permintaan dan mengimport perpustakaan *BeautifulSoup* kedalam program python, lalu proses scraping dengan memasukan url halaman yang akan dilakukan proses *scraping data,* lalu fungsi *find*() untuk dan *find_all*() untuk mencari data yang kita butuhkan. Misal kita menggunakan *find*() untuk mencari *div container* dan kita menggunakan *find_all*() untuk mencari seluruh *content* yang kita butuhkan didalam *div container* tersebut. Lalu kita hapus kalimat yang tidak kita gunakan didalam data kita seperti *"ADVERTISEMENT"*, *"SCROLL TO CONTINUE WITH CONTENT"* dll. Lalu kita masukan data kedalam file csv yang kita buat. Lalu data yang sudah dikumpulkan dilabeli sesuai kebutuhan penelitian [7].

Pada penelitian ini dataset yang digunakan merupakan berita mengenai PEMILU 2024 yang ada di website Turn Back Hoax. Dataset yang digunakan berisi berita benar dan berita salah. Adapun data berita benar dan berita salah diambil dari website Turn Back Hoax yang telah dipercaya memverifikasi berita benar dan berita salah.

2.3 Data Preparation

Pada tahapan ini data yang sudah dikumpulkan akan dipersiapkan dan dibersihkan untuk pengolahan data. Dataset yang didapatkan dari website turnbackhoax.id dengan jumlah total 2075 berita. dapat dilihat pada tabel 1 jumlah dataset berita salah berjumlah 1037 dan berita benar berjumlah 1038 berita.

Table 1 Jumlah Total Data

No	Kategori	Jumlah
1	Berita Salah	1037
2	Berita Benar	1038
	Total	**2075**

Pengolahan data program dilakukan dengan menggunakan Atribut isi berita dan label dari berita, hal ini dimaksudkan untuk memfokuskan program dalam mempelajari karakteristik atau pola berita tiap jenis.

Pengujian program dilakukan dengan membagi data latih sebanyak 80% dan data uji sebanyak 20%. Dan dilakukan pengujian tambahan untuk uji website dengan 100 data uji mengenai PEMILU 2024 yang diambil dari bulan Februari hingga bulan April. Untuk berita salah dan berita benar berjumlah 50 berita.

Dengan dataset berita yang diambil hanya mengenai PEMILU 2024. menjadikan program hanya baik dalam mengklasifikasikan berita mengenai PEMILU.

Dalam penelitian ini pengolahan data menggunakan bahasa pemrograman Python. Berikut tahapan preparation atau preprocessing data, diantaranya:

a. Labeling

Pada tahap ini dataset yang didapat akan dilabeli dengan label 0 = berita salah dan 1 = berita benar sesuai dari sumber situs *Turn Back Hoax*. Pelabelan berfungsi untuk membantu sistem dalam proses pembelajaran mesin agar sistem dapat memepelajari kriteria, ciri-ciri dan perbedaan dari berita benar dan berita salah.

b. Data Preprocessing

Pada tahap ini akan dilakukan langkah-langkah untuk membersihkan data teks lalu data teks dipersiapkan untuk pengolahan nantinya. langkah-langkah yang akan dilakukan pada proses ini adalah sebagai berikut:

- Langkah pertama adalah *Case Folding* pada tahapan ini teks di dataset akan dilakukan beberapa tahapan oleh program untuk membantu proses pengolahan data dengan menyeragamkan bentuk teks pada dataset. Diantaranya adalah Mengubah menjadi huruf kecil, Menghapus hyperlink, Menghapus tanda koma, Menghapus angka, Menghapus semua karakter yang bukan huruf dan spasi.

- Langkah kedua adalah Normalisasi Teks, Pada tahapan ini kata-kata didalam dataset akan disaring dari kata-kata singkatan menjadi kata-kata yang lebih lengkap menggunakan kalimat yang sudah dimasukan kedalam dataset normalisasi teks yang telah di buat. Proses ini dilakukan dengan memecah kata dan menyaring kata yang sesuai dengan yang ada di kolom singkat lalu diganti dengan kata yang ada di kolom hasil sehingga dapat dihasilkan kata yang lebih baku.

- Langkah ketiga adalah *Stopword Removal*, tahapan ini menggunakan data didalam daftar stopwords indonesia yang mana akan menyaring kata-kata umum yang sering muncul dalam kalimat tapi tidak memberikan informasi yang penting mengenai kalimat itu sendiri. Proses ini

dilakukan dengan memecah kata dan menghilangkan kata didalam data yang terdaftar dalam stopwords indonesia.

- Langkah terakhir adalah *Stemming*, tahapan ini menggunakan modul stemmer dari library sastrawi untuk mengubah teks ke bentuk dasarnya tanpa imbuhan dari kata sebelumnya, sehingga dapat meningkatkan hasil dari proses analisis teks [8].

c. Pembobotan TF-IDF

Pada tahap ini akan diteliti seberapa sering suatu kata muncul dalam suatu dokumen. Metode yang digunakan untuk mengetahuinya adalah dengan pembobotan TF-IDF (*Term Frequency - Invers Document Frequency*). Tujuan dari TF-IDF ada 2 yaitu yaitu:

- TF (*Term Frequency*): menghitung seberapa sering suatu kata/token yang muncul dalam dokumen. Rumus TF sesuai dengan persamaan 1 berikut ini:

$$TF(t_n, d_n) = f(t_n, d_n) \quad (1)$$

Jadi f (t_n , d_n) mendefinisikan jumlah kemunculan term-*n* pada sebuah dokumen-*n* [9].

- IDF (*Invers Document Frequency*): menilai seberapa penting suatu kata/token. Rumus dari IDF sesuai dengan persamaan 2 berikut ini:

$$IDF\ (t_n) = \log \frac{D}{df(t)} \quad (2)$$

Jadi log D/df(t) Menjelaskan jumlah dokumen pada dataset-*D* akan dibagi dengan jumlah dokumen yang mengandung term-*df(t).*

Lalu menghitung keseluruhan nilai TF-IDF dengan mengkalikan nilai keduanya. Rumus sesuai dengan persamaan 3 berikut:

$$W_{td} = TF(t_n, d_n) \times IDF\ (t_n) \quad (3)$$

Keterangan:
- d : Dokumen ke-n.
- t : Term atau kata ke-n dari kata yang di cari.
- W : Bobot kata-t terhadap dokumen ke-d.
- TF : Frekuensi/jumlah Sebuah kata dalam satu dokumen.
- IDF : *Invers Documen Frequency.*
- D : Total Dokumen.
- Df : Jumlah dokumen yang mengandung Term/Kata yang di cari [10]

d. Seleksi Fitur Chi-Square

Pada tahap ini dilakukan penyeleksian fitur yang didapat dari proses pembobotan Tf-Idf menggunakan metode *Chi-Square*, proses ini berguna untuk menghapus fitur yang kurang berguna dalam proses klasifikasi, dengan mengurutkan fitur yang memiliki nilai *Chi-Square* yang tertinggi hingga terendah, lalu fitur diseleksi dengan jumlah 3000 fitur sehingga didapatkan fitur terbaik yang dapat mendukung proses klasifikasi berita. Adapun persamaan dari pencarian fitur tertinggi *Chi-Square* sesuai dengan persamaan 4 berikut:

$$X^2 = \sum_{i=1}^{k} \frac{(f_0 - f_2)^2}{f_h} \quad (4)$$

Keterangan:
- X^2 : Chi-kuadrat.
- f_0 : Frekuensi yang diamati
- f_2 : Frekuensi yang diharapkan [11]

2.4 Modeling

Pada tahap ini data dibagi menjadi *data training* dan *data testing*, *data training* merupakan data yang digunakan untuk mempelajari pola dan karakteristik data, sedangkan *data testing* merupakan data yang digunakan untuk menguji model yang telah mempelajari pola dan karakteristik data. lalu dibangun model untuk sistem yang telah dirancang untuk mengklasifikasi berita benar dan berita salah pada berita pemilu. Penelitian ini menggunakan metode klasifikasi *Naïve Bayes*. Adapun Persamaan *Naïve Bayes* sesuai dengan persamaan 5 berikut:

$$P(x|y) = \frac{p(y|x)p(x)}{p(y)} \quad (5)$$

Keterangan:
- y : Data dengan kelas yang belum di ketahui.
- x : Hipotesis data y merupakan suatu klas spesifik.
- P(x|y) : Probalitas hipotesis x berdasarkan kondisi y (posteriori probability).
- P(x) : Probabilitas hipotesis x (prior probability).
- P(y|x) : Probabilitas y berdasarkan kondisi pada hipotesis x.
- P(y) : Probalitas dari y.

2.5 Evaluation

Pada tahap ini setelah sistem telah berhasil berjalan, maka akan dilakukan pengujian terhadap performa dan efektifitas dari sistem menggunakan *K-fold Cross Validation* dan *Confusion Matrix*.

a. K-fold Cross Validation

K-fold Cross Validation Menggunakan data Latihan (*Train*) yang dibagi menjadi subset (*Fold*) yang sama ukurannya. Lalu selama proses validasi satu subset digunakan sebagai data validasi, sementara subset sisanya digunakan sebagai data latih. Hasil semua subset diambil rata-ratanya untuk mengetahui akurasi kinerja model secara keseluruhan. Pada Gambar 2

P-ISSN : 2089-3353
E-ISSN : 2808-9162

menunjukan cara kerja dari metode klasifikasi *Cross Valiation* [12].

Fold = 1	1	2	3	4	5	6	7	8	9	10
Fold = 2	1	2	3	4	5	6	7	8	9	10
:										
Fold = 9	1	2	3	4	5	6	7	8	9	10
Fold = 10	1	2	3	4	5	6	7	8	9	10

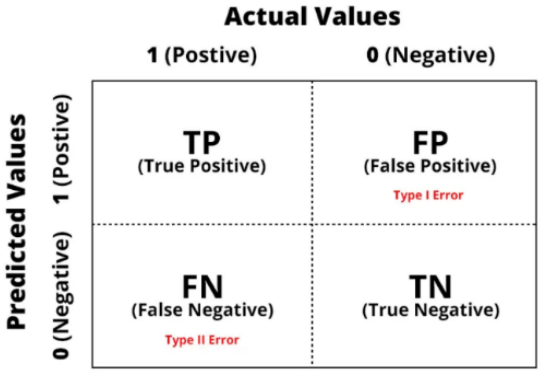
Keterangan :

Data Training
Data Testing

Gambar 2 Simulasi Cross Validation Fold=10

b. Confusion Matrix

Confusion Matrix merupakan metode yang digunakan untuk mengevaluasi model algoritma klasifikasi, sehingga pengembang dapat mengetahui sejauh mana sistem klasifikasi berjalan dengan baik. Pada gambar 3 menunjukan komponen utama dari *Confusion Matrix* sebagai berikut:



Gambar 3 Confusion Matrix

True Positive (TP) : Merupakan data Positif yang di prediksi benar.
True Negative (TN) : Merupakan data Positif yang di prediksi benar.
False Positive (FP) : Merupakan data Negatif namun diprediksi sebagai data Positif.
False Negative (FN) : Merupakan data Positif namun diprediksi sebagai data Negative.

Evaluasi klasifikasi berita dilakukan dengan menggunakan nilai akurasi. Akurasi adalah persentase jumlah data yang diprediksi dengan benar terhadap jumlah keseluruhan data. Rumus untuk menentukan akurasi sesuai dengan persamaan 6 dibawah ini [13].

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

2.6 Deployment

Dalam tahapan ini sistem akan diimplementasikan menjadi website yang dapat mengklasifikasikan berita PEMILU kedalam berita benar dan berita salah, serta dapat diakses untuk kepentingan semua orang agar mempermudah dalam klasifikasi berita yang tersebar pada masa pemilu ini, sehingga dapat diketahui berita yang benar dan berita yang salah.

3. Hasil dan Pembahasan

3.1 Data Preparation

Pada proses ini dilakukan pemrosesan pada data teks untuk dibersihkan dan diseragamkan sehingga dapat memaksimalkan klasifikasi program kepada hasil yang diinginkan. pada proses ini dilakukan beberapa pemrosesan pada data teks yang sebagai berikut :

a. Case Folding

Pada proses ini data teks diseragamkan menjadi huruf kecil dan dibersihkan dengan beberapa pengaturan perubahan yang dibutuhkan untuk memaksimalkan proses klasifikasi dari program. Perubahan yang dilakukan adalah Mengubah isi dataset menjadi huruf kecil, Menghapus hyperlink, Menghapus tanda koma, Menghapus angka, Menghapus semua karakter yang bukan huruf dan spasi. Tabel 2 menunjukan hasil perubahan pada data mentah hingga perubahan setelah dilakukan *Case Folding*.

Table 2 Hasil Case Folding

Data Mentah	Case Folding
PEMERINTAH menemukan byk org Pendukung dri kedua kubu, dalam aksi KAMPANYE yg dilakukan oleh KOALISI 02, yg berefek ricuh dan membuat Kekacauan.	pemerintah menemukan byk org pendukung dri kedua kubu dalam aksi kampanye yg dilakukan oleh koalisi yg berefek ricuh dan membuat kekacauan

b. Normalisasi Teks

Pada tahap ini data teks dilakukan proses normalisasi teks dengan mengubah kata singkatan, asing atau yang belum jelas ke kata yang lebih jelas/baku dengan data kata normalisasi yang telah disiapkan didalam dataset normalisasi teks. Tabel 3 menunjukan contoh isi dari data normalisasi teks.

Table 3 Contoh Isi Data Normalisasi

Singkat	Hasil
Abis	Habis
Byk	Banyak
car	Mobil

Dengan proses normalisasi teks yang telah dilakukan maka dapat dilihat hasil dari Normalisasi Teks pada tabel 4 dibawah ini.

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Table 4 Hasil Normalisasi Teks

Case Folding	Normalisasi Teks
pemerintah menemukan byk org pendukung dri kedua kubu dalam aksi kampanye yg dilakukan oleh koalisi yg berefek ricuh dan membuat kekacauan	pemerintah menemukan banyak orang pendukung dari kedua kubu dalam aksi kampanye yang dilakukan oleh koalisi yang berefek ricuh dan membuat kekacauan

c. Filtering (Stopword Removal)

Pada tahapan ini data teks disaring dari kata-kata yang yang kurang berguna pada proses klasifikasi program menggunakan library stopword bahasa indonesia yang ada pada library NLTK (Natural Language Tool Kit). Setelah proses *filtering* data yang kurang berguna maka didapatkan hasil dari *filtering* pada tabel 5 dibawah ini.

Table 5 Hasil Filtering (Stopword)

Normalisasi Teks	Filtering
pemerintah menemukan banyak orang pendukung dari kedua kubu dalam aksi kampanye yang dilakukan oleh koalisi yang berefek ricuh dan membuat kekacauan	pemerintah menemukan orang pendukung kubu aksi kampanye koalisi berefek ricuh kekacauan

d. Stemming

Pada tahapan ini setiap kata dalam data teks diubah ke bentuk dasar dari kata tersebut menggunakan library sastrawi. Setelah proses *stemming* didapatkan hasil dari proses *stemming* pada tabel 6 dibawah ini.

Table 6 Hasil Stemming

Filtering	Stemming
pemerintah menemukan orang pendukung kubu aksi kampanye koalisi berefek ricuh kekacauan	perintah temu orang dukung kubu aksi kampanye koalisi efek ricuh kacau

e. Pembobotan TF-IDF

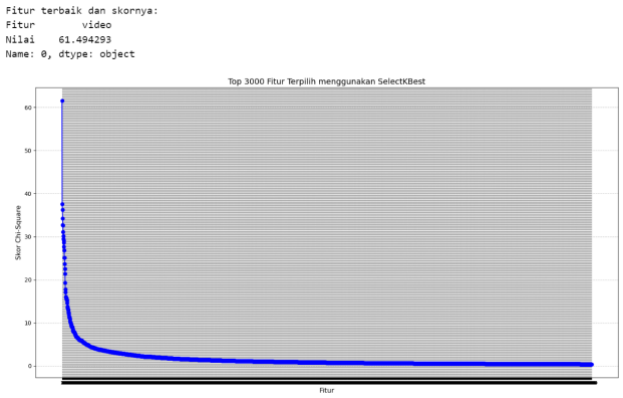
Pada tahapan ini, dilakukan proses pemisahan kolom feature dan target. Untuk teks yang sudah dipreprocessing, dimasukan kedalam kolom *Feature*(X), lalu untuk untuk label, dimasukan kedalam kolom *Target*(Y).

Lalu dibangun sebuah teknik pembobotan dengan menggunakan teknik TF-IDF untuk melihat seberapa penting suatu teks dalam kalimat tersebut. Pada tahap ini dilakukan pengimportan metode Tfidf Vectorizer dari *library* scikit-learn untuk mengubah kalimat menjadi *vector*, setelah itu akan digunakan N-gram (1,1) Unigram untuk menentukan rentang N-gramnya.

Setelah itu model Tfidf Vectorizer dilatih dengan kumpulan data teks yang telah dilakukan *text preprocessing*, proses ini akan melatih program menghitung frekuensi kata dan bobot TF-IDF untuk setiap kata dalam data teks. lalu selanjutnya dilakukan fungsi Transform untuk mengubah kumpulan data menjadi bentuk vektor Tf-Idf. Setelah itu data ditampilkan kedalam bentuk tabel dari Tf-Idf tersebut dan didapatkan 17591 kata yang terdapat dalam keseluruhan dataset.

f. Seleksi Fitur Chi-Square

Pada tahapan ini dilakukan pemilihan fitur/kata yang memiliki bobot terbaik sebanyak 3000 kata menggunakan *select feature Kbest* dan uji *Chi-Square*, kata yang diperoleh dari transformasi Tf-Idf berjumlah 17591. pemilihan fitur ini berguna untuk meningkatkan kinerja program dengan mengurangi jumlah data dari data yang kurang berguna dan fokus ke data yang berbobot saja, sehingga mengurangi resiko *overfitting* dan mempercepat pemrosesan data dan mengoptimalkan proses pelatihan program. Gambar 4 menunjukan hasil dari seleksi fitur menggunakan Kbest.



Gambar 4 seleksi fitur KBest

3.2 Modeling

Pada tahapan ini dilakukan pembangunan model sistem yang telah dirancang untuk mengklasifikasi berita benar dan berita salah pada masa PEMILU, proses ini akan dilakukan dengan beberapa tahapan sebagai berikut:

a. Train & Test Split

Pada proses ini data akan dibagi menjadi data latih dan data uji. Data latih digunakan untuk mempelajari bentuk dari data benar dan data salah yang akan diklasifikasi, proses ini menggunakan 80% data total dataset yang yang telah diproses, data uji digunakan untuk menguji program yang telah melakukan latihan dengan data benar dan data salah sehingga dapat dijadikan untuk bahan evaluasi, proses ini menggunakan 20% data dari total dataset yang telah diproses.

P-ISSN : 2089-3353
E-ISSN : 2808-9162

b. Proses Modeling

Pada proses ini digunakan sebuah berita untuk diklasifikasi kedalam berita benar dan berita salah menggunakan metode klasifikasi *Naïve Bayes*. Data teks yang telah dimasukan, diolah dengan proses *Text Processing* yang telah dibuat lalu diklasifikasi dengan metode klasifikasi *Naïve Bayes* seetelah itu akan dimunculkan hasil klasifikasinya.

3.3 Evaluation

Pada tahapan ini hasil dari program akan dievaluasi performa metode klasifikasi yang digunakan menggunakan *Confusion Matrix* dan *Cross Validation*, proses ini dilakukan untuk mengetahui seberapa baik program bekerja dalam mengklasifikasi berita PEMILU. Adapun nilai yang diukur dalam penelitian ini adalah nilai *Accuracy*, *Precision*, *Recall*, dan *F1-Score*.

a. Hasil Evaluasi Cross Validation

Tahap ini menunjukan hasil performa metode klasifikasi *Naïve Bayes* yang dihitung menggunakan *Cross Validation* sebanyak 10-fold, proses evaluasi *Cross Validation* akan dilakukan 10 kali lalu dilihat rata-rata dari hasil evaluasi *Cross Validation* tersebut. Pada tabel 7 menunjukan hasil dari evaluasi dari *Cross Validation.*

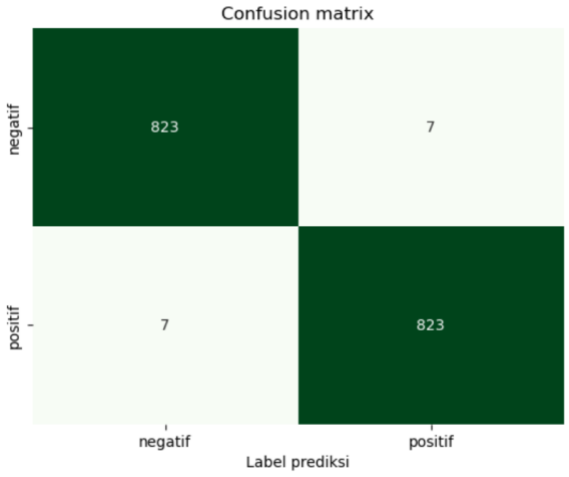
Table 7 Hasil Cross Validation

Fold	Hasil
1	0.99
2	0.98
3	1
4	0.98
5	0.99
6	0.98
7	0.99
8	0.99
9	0.98
10	0.98
Rata-rata	**0.9**

Dari hasil yang didapatkan dapat dilihat bahwa akurasi metode klasifikasi *Naïve Bayes* sudah bagus dengan rata-rata akurasi 0.9. sehingga dapat disimpulkan dari hasil evaluasi *cross validation* diatas bahwa metode klasifikasi *Naïve Bayes* cocok untuk dijadikan metode klasifikasi berita benar dan berita salah.

b. Hasil Evaluasi Confusion Matrix

Pada tahap ini ditunjukan hasil klasifikasi berita dari metode *Naïve Bayes* menggunakan *Confusion Matrix.* Metode ini menunjukan berapa banyak data yang diklasifikasi dengan benar dan berapa banyak data yang diklasifikasi dengan salah, lalu hasil evaluasi dari klasifikasi akan dijumlahkan dan dapat diketahui berapa nilai dari *Accuracy*-nya. Adapun hasil dari proses evaluasi *Confusion Matrix* adalah sebagai berikut:



Gambar 5 Hasil *Confusion Matrix*

Hasil dari *Confusion Matrix* pada gambar 5 menunjukan metode *Naïve Bayes* berhasil mengklasifikasi data dengan nilai *True Positive* (TP) = 823, *True Negative* (TN) = 823, *False Positive* (FP) = 7, *False Negative* (FN) = 7.

```
test time: 0.014s
accuracy: 0.992
              precision    recall  f1-score   support

     negatif       0.99      0.99      0.99       830
     positif       0.99      0.99      0.99       830

    accuracy                           0.99      1660
   macro avg       0.99      0.99      0.99      1660
weighted avg       0.99      0.99      0.99      1660


Text(0.5, 23.52222222222222, 'Label prediksi')
```

Gambar 6 hasil *Accuracy Confusion Matrix*

Setalah itu diketahui hasil *Accuracy* pada gambar 6 menunjukan bahwa kinerja metode klasifikasi *Naïve Bayes* bekerja dengan baik dengan rata-rata performa *Naïve Bayes* 0.99, dan dapat disimpulkan bahwa metode klasifikasi ini bagus untuk digunakan dalam klasifikasi berita benar dan berita salah.

3.4 Deployment

a. Save and Read Pickle File

Pada tahapan ini syntax dan data yang dibutuhkan untuk diimplementasikan kedalam website akan disimpan kedalam file pickle dengan format sav, lalu dimasukan kedalam program website, adapun syntax dan data yang diambil dari program Python adalah sebagai berikut:

P-ISSN : 2089-3353
E-ISSN : 2808-9162

- Fitur Tf-Idf yang telah dipilih menggunakan metode *Chi-Square* sebanyak 3000 fitur.
- Metode klasifikasi *Naïve Bayes*.

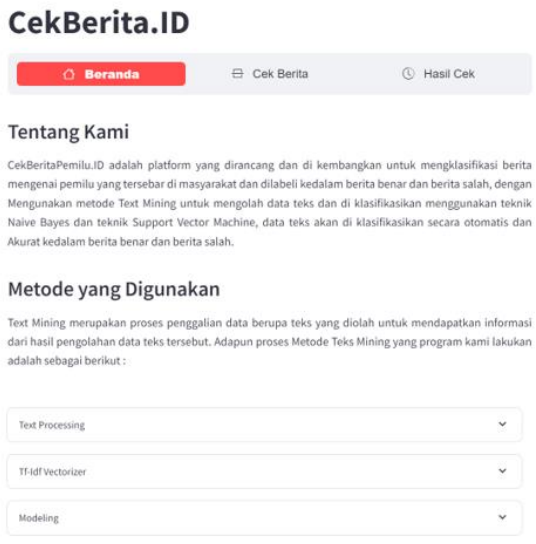
b. Building Application

Pada tahapan ini dikembangkan website klasifikasi berita PEMILU yang akan mengklasifikasikan berita kedalam berita benar dan berita salah menggunakan metode klasifikasi *Naïve Bayes* yang sudah dibuat modelnya sebelumnya didalam program python, lalu hasil dari klasifikasi akan ditampilkan didalam halaman website.

Penelitian ini menggunakan streamlit untuk pengembangan websitenya, dimana streamlit merupakan perangkat lunak pengembangan website menggunakan bahasa python. sehingga, dapat mengimplementasikan program python yang telah dibuat menjadi website.

Untuk mengimplementasikan program Python menjadi website, dibuat sebuah program python baru untuk pengembangan website, lalu syntax yang sudah disimpan didalam file sav, dimasukan kedalam program website untuk diimplementasikan syntaxnya sebagai fungsi dari website. lalu menghubungkan website dengan google spreadsheet untuk menjadi database website, sehingga hasil klasifikasi yang telah dilakukan akan terekam kedalam database, sehingga dapat dipanggil kembali agar dapat ditampilkan didalam website.

Bagian dalam website terdiri dari beranda, cek berita dan hasil cek, tiap bagian memiliki fungsinya masing-masing yang digunakan untuk membantu proses klasifikasi berita PEMILU.

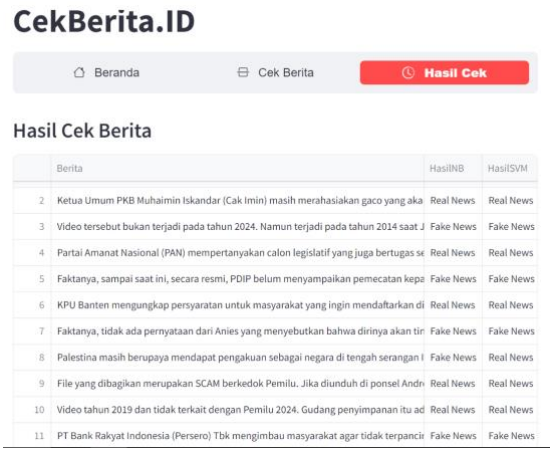


Gambar 7 Halaman Beranda Website

Pada gambar 7 menunjukan halaman beranda menjelaskan latar belakang website buat, metode yang digunakan dan teknik yang digunakan untuk menjelaskan fungsi dari website dan bagamana website ini bekerja.

Gambar 8 Halaman Cek Berita Website

Lalu dapat dilihat pada gambar 8 diatas merupakan halaman cek berita disediakan layanan untuk klasifikasi berita PEMILU kedalam berita benar dan berita salah.



Gambar 9 Halaman Hasil Cek

Lalu pada gambar 9 menunjukan halaman hasil cek menampilkan hasil dari klasifikasi berita PEMILU sebelumnya sehingga dapat dilihat kembali hasil klasifikasi berita terdahulu. Setelah selesai mengembangakan program website menggunakan streamlit, program website diupload kedalam Github lalu dijalankan menggunakan hosting dari streamlit community cloud agar dapat diakses diinternet.

c. Uji Hasil Klasifikasi

Setelah selesai mengembangkan website klasifikasi berita PEMILU, website diuji untuk klasifikasi menggunakan 100 data uji dari berita benar dan berita salah yang diambil dari website Turn Back Hoax dan hasilnya dapat dilihat pada table 8 dibawah ini.

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Table 8 Hasil Uji Klasifikasi

Berita	Diklasifikasi Benar	Diklasifikasi Salah
Salah	44	6
Benar	47	3
Total	**91**	**9**

Dapat dilihat pada tabel 10 bahwa hasil klasifikasi dari website sudah optimal dengan hasil klasifikasi berita salah yang diklasifikasi benar sebanyak 44 berita dan yang diklasifikasi salah sebanyak 6 berita, sehingga tingkat akurasi klasifikasi untuk berita salah sebesar 88%. Lalu, untuk berita benar yang diklasifikasi benar sebanyak 47 berita dan yang diklasifikasi salah sebanyak 3 berita, sehingga tingkat akurasi klasifikasi untuk berita benar sebesar 91%.

Dan didapatkan total akurasi klasifikasi untuk keseluruhan berita yang diklasifikasi benar sebanyak 91 berita dan yang diklasifikasi salah sebanyak 9 berita, sehingga didapatkan hasil tingkat akurasi klasifikasi metode *Naive Bayes* untuk berita PEMILU 2024 sebesar 91%. Hasil dari klasifikasi yang didapat didukung oleh bentuk data yang baik dan proses pengolahan data untuk proses latihan klasifikasi sudah baik pula. Jadi, dapat disimpulkan bahwa metode ini cocok untuk dijadikan metode klasifikasi data teks secara umum terutama dalam klasifikasi berita hoax.

Table 9 Perbandingan penelitian sebelumnya

Judul	Metode	Akurasi	Perbandingan
Analisis Data Mining Klasifikasi Berita salah COVID 19 Menggunakan Algoritma Naive Bayes.	Naïve Bayes	86.30%	penelitian ini tidak menggunakan Seleksi Fitur Chi-Square
Sistem prediksi berita palsu tentang virus covid-19 menggunakan algoritma support vector machine (svm).	Support Vector Machine (SVM).	78%	Penelitian ini tidak menggunakan Algoritma Naïve bayes dan Seleksi Fitur Chi-Square

Tabel 9 menunjukan perbandingan dengan penelitian sebelumnya. Hasil diatas menunjukan perbandingan akurasi dan perbedaan model penelitian dengan penelitian sebelumnya. Dimana penelitian ini menggunakan metode Naïve Bayes dan seleksi fitur menggunakan Chi-Square.

4. Kesimpulan

4.1 Kesimpulan

Pada penelitian ini dapat ditarik kesimpulan bahwa metode *Naïve Bayes* bagus untuk digunakan dalam klasifikasi berita PEMILU, sehingga dapat digunakan dalam kasus klasifikasi data teks berita dan dapat dicoba untuk klasifikasi data teks lainnya, seleksi fitur menggunakan metode *Chi-Square* dalam pemilihan fitur terbaik untuk latihan model klasifikasi didapatkan hasil evaluasi akurasi yang cukup baik dan optimal. Metode klasifikasi *Naïve Bayes* mendapatkan nilai akurasi sebesar 91% hasil tersebut di dukung dengan data yang sudah baik. Berdasarkan hasil pengujian tersebut, diharapkan hasil penelitian ini dapat memberikan sumbangan bagi literasi digital masyarakat mengenai kearuratan berita pemilu.

4.2 Saran

Untuk peneliti selanjutnya diharapkan dapat mengembangkan kekurangan dari penelitian ini dengan mencoba memasukan dataset berita dari berbagai macam sumber sehingga hasil klasifikasi dapat menjadi lebih akurat untuk klasifikasi berita salah dan mengembangkan website dengan menggunakan perangkat lunak yang lebih mudah dan baik dalam menjalankan website untuk *Data Mining* atau fungsi dari website yang kurang dalam penelitian ini.

Hasil penelitian ini diharapkan dapat terus di kembangkan dan dapat menjadi rujukan bagi peneliti selanjutnya untuk kegunaan penelitian klasifikasi berita salah terutama dalam implementasi program website, sehingga penelitian ini dapat bermanfaat untuk masyarakat secara luas dalam pemberantasan berita salah yang dapat merugikan rakyat.

Daftar Rujukan

[1] A. M. Kaplan and M. Haenlein, “Users of the world , unite ! The challenges and opportunities of Social Media,” no. December, 2017, doi: 10.1016/j.bushor.2009.09.003.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake News Detection on Social Media : A Data Mining Perspective Fake News Detection on Social Media : A Data Mining Perspective,” no. August, 2017, doi: 10.1145/3137597.3137600.

[3] B. H. K. Kominfo, “Menkominfo: Isu Hoaks Pemilu Meningkat Hampir 10 Kali Lipat,” 2023.

[4] F. Prasetya, “Analisis Data Mining Klasifikasi Berita Hoax COVID 19 Menggunakan Algoritma Naive Bayes,” vol. 4, no. September, pp. 132–139, 2022, doi: 10.30865/json.v4i1.4852.

[5] M. F. Muttaqin, T. Bukhori, Y. Yanto, N. Agustina, and M. Naseer, “Sistem Prediksi Berita Palsu Tentang Virus Covid-19 Menggunakan Algoritma Support Vector Machine (Svm),” *Naratif J. Nas. Riset, Apl. dan Tek. Inform.*, vol. 5, no. 1, pp. 26–33, 2023, doi:

P-ISSN : 2089-3353
E-ISSN : 2808-9162

10.53580/naratif.v5i1.187.

[6] D. Rahma Putri, B. Arif Dermawan, I. Purnamasari, U. H. Singaperbangsa Karawang Jl Ronggo Waluyo, and T. Timur, "Implementasi Modified Enhanced Confix Stripping Stemmer pada Klasifikasi Fake News Covid-19," *J. Sains Komput. Inform. (J-SAKTI*, vol. 5, no. 2, pp. 589–600, 2021.

[7] A. Cardova and A. Hermawan, "Implementasi Metode LSTM Untuk Mengklasifikasi Berita Palsu Pada PolitiFact," vol. 13, no. 3, pp. 471–479, 2023.

[8] J. Sanjaya, B. Priyatna, and S. S. Hilabi, "Analisis Sentimen Terhadap Opini Proyek Kereta Cepat Menggunakan Metode Naïve Bayes Classifier," vol. 14, no. 1, pp. 263–270, 2024.

[9] C. H. Yutika, A. Adiwijaya, and S. Al Faraby, "Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 422, 2021, doi: 10.30865/mib.v5i2.2845.

[10] D. Muriyatmoko, Taufiqurrahman, and A. Humam, "Analisis Sentimen Masyarakat Terhadap Konflik Rusia dan Ukraina Menggunakan Metode Naïve Bayes pada Media Sosial Twitter," *Metik J.*, vol. 6, no. 2, pp. 140–145, 2022, doi: 10.47002/metik.v6i2.375.

[11] R. Aziz, T. M. Fahrudin, W. Syaifullah, and J. Saputra, "Analisis Sentimen Kepuasan Pengguna OYO Di Playstore Dengan Multinomial Naive Bayes dan Chi-square," vol. 14, no. 1, pp. 166–175, 2024.

[12] H. Hafid, "Penerapan K-Fold Cross Validation untuk Menganalisis Kinerja Algoritma K-Nearest Neighbor pada Data Kasus Covid-19 di Indonesia," *J. Math.*, vol. 6, no. 2, pp. 161–168, 2023.

[13] A. Y. A. Nugraha and F. F. Abdulloh, "Optimasi Naive Bayes dan Cosine Similarity Menggunakan Particle Swarm Optimization Pada Klasifikasi Hoax Berbahasa Indonesia," *J. Media Inform. Budidarma*, vol. 6, no. 3, p. 1444, 2022, doi: 10.30865/mib.v6i3.4170.