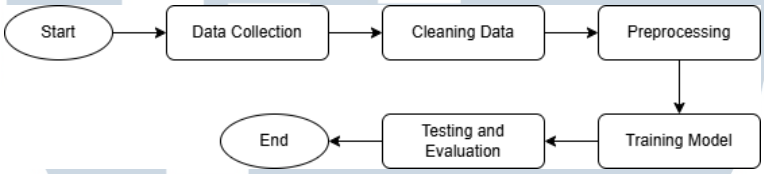
BAB 3

METODOLOGI PENELITIAN

Metodologi penelitian ini dirancang untuk memberikan gambaran terperinci mengenai tahapan yang dilakukan dalam mendeteksi berita hoax menggunakan algoritma Naive Bayes. Setiap tahap dimulai dari pengumpulan data hingga evaluasi model, yang disusun secara sistematis untuk memastikan keakuratan dan validitas hasil penelitian.



Gambar 3.1. Flowchart Research Metodology

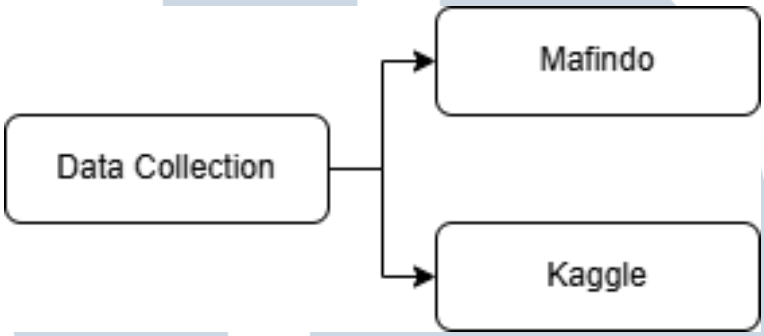
Gambar 3.1 menunjukkan alur penelitian yang terdiri dari lima tahapan utama dalam proses deteksi berita hoax menggunakan algoritma Naive Bayes. Proses dimulai dari tahap Data Collection, di mana data berupa berita hoax dan non-hoax dikumpulkan dari berbagai sumber terpercaya. Selanjutnya, data yang telah dikumpulkan melewati tahap Cleaning Data, yaitu proses membersihkan data dari elemen yang tidak relevan, seperti tanda baca, simbol, dan duplikasi data.

Tahap berikutnya adalah Preprocessing, di mana data diproses lebih lanjut melalui langkah-langkah seperti tokenisasi, penghapusan stopwords, dan stemming, sehingga menghasilkan representasi teks yang siap digunakan untuk pelatihan model. Data yang telah diproses kemudian digunakan dalam tahap Training Model, di mana algoritma Naive Bayes dilatih menggunakan teknik ekstraksi fitur seperti Count Vectorizer dan TF-IDF Vectorizer dengan variasi N-gram.

Setelah model selesai dilatih, dilakukan tahap Testing and Evaluation untuk menguji kinerja model menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score. Proses ini bertujuan untuk menentukan efektivitas dan efisiensi model dalam mendeteksi berita hoax. Akhir dari alur penelitian ditandai dengan analisis hasil evaluasi, yang menjadi dasar untuk menarik kesimpulan dan memberikan rekomendasi.

3.1 Pengumpulan Data

Tahap pertama dalam penelitian ini adalah pengumpulan data dari dua sumber utama, yaitu API Mafindo dan dataset dari Kaggle Gambar 3.2.



Gambar 3.2. Data Collection

Dataset pertama diperoleh dari API Mafindo, yang merupakan organisasi anti-hoax terpercaya di Indonesia. Data diambil menggunakan modul `requests` pada Python melalui endpoint https://yudistira.turnbackhoax.id/api/antihoax/. Proses pengambilan data menghasilkan 14.552 entri, di mana 14.057 entri dilabeli sebagai hoax dan 495 entri dilabeli sebagai berita fakta.

Karena terdapat ketidakseimbangan data antara kelas hoax dan fakta, dataset kedua ditambahkan dari sumber Kaggle dengan judul ”Indonesia News Dataset (2024)” [26]. Dataset ini terdiri dari artikel-artikel berita yang diambil dari sumber terpercaya seperti Kompas.com, Tempo.co, dan Detik.com, mencakup periode Januari hingga September 2024. Dataset ini dipilih untuk melengkapi data fakta yang sebelumnya berjumlah sedikit dan untuk memastikan representasi yang memadai dari berbagai jenis berita.

Tabel 3.1. Data Mafindo & Kaggle

	Hoax	**Fakta**
Mafindo	14.057	459
Kaggle	0	45.234
Total	14.057	45.693

Setelah proses penggabungan, dataset akhir terdiri dari jumlah yang seimbang antara berita hoax dan berita fakta. Hal ini dilakukan untuk memastikan performa model yang optimal, tanpa bias terhadap salah satu kelas.

Dataset ini selanjutnya diproses lebih lanjut pada tahap preprocessing untuk digunakan dalam pelatihan dan pengujian model.

3.2 Pembersihan Data

Proses pembersihan data dilakukan untuk menghilangkan elemen-elemen yang tidak relevan dan meningkatkan kualitas dataset sebelum digunakan dalam tahap preprocessing.

Pada dataset hoax yang diambil dari API Mafindo, banyak entri mengandung metadata atau label status seperti ”Klarifikasi,” ”Berita,” ”Dalam Proses,” dan ”Edukasi,” yang tidak secara langsung relevan dengan tugas klasifikasi. Label-label ini dihapus secara sistematis menggunakan filter otomatis. Selain itu, dilakukan penghapusan elemen seperti karakter khusus, URL, tautan web, spasi berlebih, dan baris baru yang dapat memperkenalkan noise ke dalam data tekstual.

Dataset fakta yang diperoleh dari Kaggle juga menjalani proses pembersihan yang serupa, tetapi disesuaikan dengan karakteristik data tersebut. Informasi pengantar seperti ”KOMPAS.com” dan tag lokasi dihapus untuk menghindari bias, serta dilakukan penghapusan karakter khusus, normalisasi teks, dan penghapusan spasi berlebih.

Proses ini memastikan bahwa kedua dataset memiliki teks yang bersih dan relevan untuk digunakan dalam tahap berikutnya. Setelah pembersihan, kedua dataset digabungkan, menghasilkan jumlah data yang seimbang antara berita hoax dan berita fakta, yang siap untuk diproses lebih lanjut pada tahap preprocessing.

3.3 Preprocessing

Setelah proses cleaning data selesai, langkah selanjutnya adalah preprocessing data untuk mempersiapkan dataset agar dapat digunakan dalam pelatihan dan pengujian model pembelajaran mesin. Dataset hoax yang diperoleh dari API Mafindo digabungkan dengan dataset berita fakta dari Kaggle setelah dilakukan pembersihan. Untuk mengatasi ketidakseimbangan jumlah data antara berita hoax dan berita fakta, dilakukan pengambilan sampel acak dari berita fakta pada dataset Kaggle, yang menjadi total berita

fakta 15.495 dan berita hoax 14.057, seperti yang ditunjukkan pada Tabel 3.2. Total data yang digunakan dalam penelitian ini adalah sebanyak 29.552 artikel. Proses penggabungan menghasilkan dataset gabungan yang hanya menyertakan dua kolom utama, yaitu kolom teks berita penuh (content) dan kolom label biner (is hoax), di mana label 0 menunjukkan berita fakta dan label 1 menunjukkan berita hoax.

Berita Fakta	**Berita Hoax**
15.495	14.057
Total: 29.552 artikel	

Tabel 3.2. Distribusi Berita fakta dan Berita Hoax

Langkah selanjutnya adalah memeriksa dan menghapus entri yang mengandung nilai hilang (missing values) serta entri duplikat untuk menjaga integritas data. Setelah itu, teks dalam dataset diproses lebih lanjut melalui berbagai tahapan pembersihan menggunakan pustaka Python seperti `re` dan `string`. Tahapan ini meliputi mengubah teks menjadi huruf kecil (lowercasing), menghapus tanda baca, angka, simbol, tag HTML, dan tautan web yang tidak relevan, serta menghilangkan karakter baris baru dan spasi berlebih. Kata-kata umum yang tidak memiliki makna signifikan, atau yang dikenal sebagai stopwords, juga dihapus menggunakan pustaka `nltk`. Untuk memastikan teks benar-benar bersih, emoji yang mungkin terdapat dalam teks dihapus menggunakan pustaka `emot`.

Selain pembersihan teks, proses stemming juga diterapkan untuk mengubah kata-kata menjadi bentuk dasarnya dengan menghilangkan imbuhan, awalan, dan akhiran. Langkah ini dilakukan menggunakan pustaka Sastrawi Stemmer, yang dirancang khusus untuk bahasa Indonesia. Proses stemming bertujuan untuk menyederhanakan variasi kata sehingga model lebih mudah mengenali pola dalam data.

Setelah pembersihan teks dan stemming selesai, data melalui proses tokenisasi, yaitu pemecahan teks menjadi unit-unit kecil seperti kata atau frasa menggunakan pustaka `nltk`. Langkah ini bertujuan untuk mempersiapkan teks agar dapat direpresentasikan dalam bentuk numerik pada tahap pelatihan model. Dataset yang telah selesai diproses kemudian dibagi menjadi dua subset dengan rasio 80:20, di mana 80% data digunakan untuk melatih model *(training data)*, dan 20% sisanya digunakan untuk menguji

kinerja model *(testing data)*. Pembagian ini ditunjukkan pada Tabel 3.3. Tahap *preprocessing* ini memastikan bahwa dataset yang digunakan bersih, konsisten, dan siap untuk digunakan dalam proses pelatihan algoritma Naive Bayes dengan teknik ekstraksi fitur yang diterapkan pada tahap berikutnya.

80%	**20%**
23.567	5.892

Tabel 3.3. Rasio Pembagian Data

3.4 Melatih Model

Pada tahap ini, data yang telah melalui proses preprocessing digunakan untuk melatih model klasifikasi berbasis algoritma Naive Bayes. Pelatihan model bertujuan untuk membangun sistem yang mampu mengenali pola dalam data teks dan mengklasifikasikan artikel berita ke dalam dua kategori: hoax atau fakta.

Dalam penelitian ini, algoritma Naive Bayes dipilih karena kemampuannya yang efisien dan sederhana untuk menangani data berbasis teks. Proses pelatihan dilakukan dengan menerapkan dua teknik ekstraksi fitur utama, yaitu Count Vectorizer dan TF-IDF Vectorizer, yang masing-masing menghasilkan representasi numerik dari data teks berdasarkan frekuensi kata dan bobot relevansi kata. Selain itu, pendekatan N-gram digunakan untuk menangkap pola hubungan antar kata dengan variasi unigram (N=1), bigram (N=2), trigram (N=3), unigram & bigram (1,2), bigram & trigram (1,3).

Pada tahap ini, algoritma Naive Bayes mempelajari distribusi probabilitas kata-kata dalam setiap kelas (hoax dan fakta) berdasarkan teknik ekstraksi fitur yang diterapkan. Proses pelatihan dilakukan secara sistematis untuk setiap kombinasi teknik ekstraksi fitur (Count Vectorizer dan TF-IDF Vectorizer) dengan variasi N-gram (unigram, bigram, trigram).

Hasil dari proses pelatihan adalah model klasifikasi yang mampu memprediksi kelas artikel berdasarkan pola yang telah dipelajari. Model ini kemudian dievaluasi pada tahap berikutnya untuk menentukan performanya menggunakan metrik seperti akurasi, presisi, recall, dan F1-score.

3.5 Testing dan Evaluasi

Setelah model klasifikasi selesai dilatih, langkah berikutnya adalah menguji dan mengevaluasi performanya menggunakan data uji. Proses pengujian bertujuan untuk menilai sejauh mana model dapat mengenali pola dalam data baru dan mengklasifikasikan artikel ke dalam kategori hoax atau fakta secara akurat. Evaluasi dilakukan untuk setiap kombinasi teknik ekstraksi fitur (Count Vectorizer dan TF-IDF Vectorizer) dengan variasi N-gram (unigram, bigram, trigram).

Proses evaluasi dilakukan menggunakan beberapa metrik utama untuk mengukur kinerja model, yaitu:

1. **Accuracy**: Mengukur persentase prediksi yang benar dari total data uji.
2. **Precision**: Menilai sejauh mana prediksi positif model benar-benar sesuai dengan kelas positif (hoax).
3. **Recall**: Mengukur kemampuan model untuk mendeteksi seluruh data yang termasuk dalam kelas positif (hoax).
4. **F1-Score**: Rata-rata harmonis antara presisi dan recall, yang memberikan penilaian seimbang terhadap kedua metrik tersebut.

Sebagai bagian dari evaluasi, confusion matrix juga digunakan untuk memberikan gambaran detail tentang distribusi prediksi model, termasuk jumlah prediksi benar dan salah untuk setiap kelas. Informasi ini membantu dalam menganalisis kekuatan dan kelemahan model, serta memahami pola kesalahan yang terjadi, seperti prediksi berita fakta sebagai hoax atau sebaliknya.

Hasil evaluasi dari setiap kombinasi teknik ekstraksi fitur dan variasi N-gram kemudian dibandingkan untuk menentukan konfigurasi terbaik yang memberikan performa tertinggi. Konfigurasi terbaik diidentifikasi berdasarkan keseimbangan antara akurasi, presisi, recall, dan F1-score, sehingga dapat digunakan untuk mendeteksi berita hoax secara efektif pada data baru.