LAPORAN PROYEK

KLASIFIKASI UJARAN KEBENCIAN (HATE SPEECH DETECTION) MENGGUNAKAN NAIVE BAYES

Disusun Untuk Memenuhi Tugas UTS Mata Kuliah Kecerdasan Buatan

Dosen Pengampu: Dr. Muhamad Fatchan, S.KOM., M.KOM.

Disusun Oleh:

Mohammad Hapiyansyah

312210243

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS PELITA BANGSA

KAB. BEKASI

2025

BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi dan komunikasi telah mengubah cara manusia berinteraksi. Media sosial seperti Twitter, Facebook, dan Instagram telah menjadi platform utama untuk berbagi informasi, berekspresi, dan berpartisipasi dalam diskusi publik. Menurut data APJII (Asosiasi Penyelenggara Jasa Internet Indonesia) tahun 2023, pengguna internet di Indonesia telah mencapai 215,63 juta jiwa atau sekitar 78,19% dari total populasi. Dari jumlah tersebut, sebagian besar menggunakan media sosial sebagai sarana komunikasi dan interaksi sehari-hari.

Namun, kebebasan berekspresi di media sosial seringkali disalahgunakan untuk menyebarkan konten negatif, termasuk ujaran kebencian (*hate speech*). Ujaran kebencian adalah komunikasi verbal, tertulis, atau perilaku yang menyerang, mengancam, atau menghina individu atau kelompok berdasarkan atribut tertentu seperti ras, agama, etnis, jenis kelamin, orientasi seksual, disabilitas, atau identitas lainnya.

Di Indonesia, fenomena ujaran kebencian di media sosial semakin marak, terutama dalam konteks politik, agama, dan isu-isu SARA (Suku, Agama, Ras, dan Antar-golongan). Dampak negatif dari penyebaran konten ini sangat signifikan; hal tersebut dapat memicu konflik sosial, meningkatkan polarisasi di masyarakat, mendorong diskriminasi, dan mengganggu stabilitas sosial.

Besarnya volume data dan kecepatan penyebaran ujaran kebencian di platform digital membuat proses moderasi konten secara manual menjadi tidak lagi memadai, lambat, dan memakan banyak sumber daya. Oleh karena itu, diperlukan sebuah sistem otomatis yang dapat mendeteksi dan menyaring konten ujaran kebencian secara cepat dan akurat.

Di sinilah peran *Natural Language Processing* (NLP) dan *Machine Learning* (ML) menjadi krusial. Dengan memanfaatkan teknik NLP dan algoritma ML, komputer dapat dilatih untuk mengenali pola-pola bahasa yang mengindikasikan ujaran kebencian. Salah satu algoritma klasifikasi teks yang populer, sederhana, namun efektif adalah **Naive Bayes**. Algoritma ini dikenal baik untuk tugas klasifikasi teks, bekerja baik dengan dataset kecil, dan cepat dalam proses

training, menjadikannya pilihan yang tepat untuk membangun model dasar (*baseline*) deteksi ujaran kebencian.

1.2 Identifikasi Masalah

Berdasarkan latar belakang yang telah diuraikan, dapat diidentifikasi beberapa permasalahan sebagai berikut:

1. Maraknya penyebaran ujaran kebencian (*hate speech*) di media sosial berbahasa Indonesia, khususnya yang berkaitan dengan isu politik dan SARA.
2. Keterbatasan moderasi konten secara manual yang tidak mampu menangani volume data yang besar dan kecepatan penyebaran konten negatif di media sosial.
3. Perlunya sistem deteksi otomatis yang cepat dan akurat untuk mengklasifikasikan teks berbahasa Indonesia sebagai ujaran kebencian atau bukan.
4. Tantangan teknis dalam membangun model klasifikasi, terutama dalam menangani dataset yang tidak seimbang (*imbalanced dataset*), di mana jumlah data ujaran kebencian jauh lebih sedikit daripada data non-ujaran kebencian.

1.3 Batasan Masalah

Untuk menjaga agar penelitian dalam proyek ini tetap fokus dan terarah, maka ditetapkan batasan masalah sebagai berikut:

1. Proyek ini berfokus pada **klasifikasi biner**, yaitu membedakan teks ke dalam dua kategori: *Hate Speech* (HS) dan *Non-Hate Speech* (Non_HS).
2. Dataset yang digunakan terbatas pada "ID Hate Speech Dataset" yang berisi 713 *tweet* berbahasa Indonesia dalam konteks Pilkada DKI Jakarta 2017.
3. Algoritma *machine learning* yang digunakan sebagai model utama adalah **Multinomial Naive Bayes**.
4. Metode *feature extraction* yang digunakan adalah **TF-IDF** (*Term Frequency-Inverse Document Frequency*) dengan kombinasi *unigram* dan *bigram* (n-gram range 1,2).
5. Proyek ini tidak mencakup proses *deployment* model ke dalam aplikasi *real-time* atau produksi.

1.4 Rumusan Masalah

Berdasarkan latar belakang, identifikasi, dan batasan masalah di atas, rumusan masalah dalam proyek ini adalah:

1. Bagaimana membangun sistem klasifikasi otomatis yang dapat mendeteksi ujaran kebencian dalam teks berbahasa Indonesia menggunakan metode Naive Bayes?
2. Bagaimana teknik *Random Under-Sampling* diterapkan untuk menangani dataset yang tidak seimbang (*imbalanced*) agar dapat meningkatkan performa model?
3. Bagaimana performa model Naive Bayes dalam mendeteksi ujaran kebencian, diukur menggunakan metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score*?
4. Apa saja kata-kata atau pola (*unigram* dan *bigram*) yang paling berpengaruh dalam mengidentifikasi ujaran kebencian pada dataset yang digunakan?

1.5 Tujuan Proyek

Proyek ini bertujuan untuk:

1. Membangun sistem klasifikasi otomatis untuk mendeteksi ujaran kebencian dalam teks berbahasa Indonesia menggunakan algoritma Naive Bayes.
2. Menerapkan teknik *Random Under-Sampling* untuk menangani masalah *imbalanced dataset*.
3. Mengevaluasi performa model klasifikasi ujaran kebencian menggunakan metrik standar (*Accuracy*, *Precision*, *Recall*, *F1-Score*).
4. Memberikan visualisasi dan interpretasi hasil analisis, termasuk kata-kata yang paling berpengaruh dalam klasifikasi.

1.6 Manfaat Proyek

Proyek ini diharapkan dapat memberikan manfaat sebagai berikut:

1. **Manfaat Akademis:** Memberikan contoh penerapan praktis konsep *Natural Language Processing* (NLP) dan *Machine Learning* (ML), khususnya algoritma Naive Bayes dan penanganan *imbalanced data*, dalam studi kasus deteksi ujaran kebencian.
2. **Manfaat Praktis:** Hasil dari proyek ini dapat digunakan sebagai model dasar (*baseline*) untuk pengembangan sistem moderasi konten otomatis pada platform media sosial atau aplikasi digital lainnya.
3. **Manfaat Sosial:** Membantu upaya mengurangi penyebaran konten negatif dan ujaran kebencian di ruang digital, sehingga berkontribusi dalam menciptakan lingkungan media sosial yang lebih aman dan positif.

1.7 Sistematika Penulisan

Struktur penulisan untuk masing-masing bab diuraikan sebagai berikut:

BAB I PENDAHULUAN

Bab ini membahas latar belakang, masalah, dan batasan, tujuan, manfaat, sistematika penulisan.

BAB II METODOLOGI

Bab ini menjelaskan secara rinci langkah-langkah teknis dan sistematis yang digunakan untuk menjawab rumusan masalah. Pembahasan mencakup kerangka alur penelitian, metode pengumpulan data, tahapan preprocessing data, teknik pelabelan, implementasi model, serta metode pengujian dan evaluasi performa model.

BAB III HASIL DAN PEMBAHASAN

Bab ini membahas sistem yang berjalan. Analisa sistem juga mencakup analisis masalah dan kebutuhan sistem, serta metode untuk menerapkan perancangan sistem yang diperlukan.

BAB IV PENUTUP

Bab terakhir dari proses penulisan laporan berisi kesimpulan dan rekomendasi dari materi yang telah dibahas pada bab-bab sebelumnya.

BAB II
METODOLOGI

2.1 Dataset

Dataset yang digunakan adalah **ID Hate Speech Dataset** yang berisi 713 tweet berbahasa Indonesia dengan distribusi:

Label	Jumlah	Persentase
Non_HS (Non Hate Speech)	608	85.27%
HS (Hate Speech)	105	14.73%

Karakteristik Dataset:

- Sumber: Tweet berbahasa Indonesia
- Konteks: Pilkada DKI Jakarta 2017
- Format: CSV dengan kolom Label dan Tweet
- Masalah: Dataset tidak seimbang (imbalanced)

2.2 Preprocessing Teks

Tahapan preprocessing yang dilakukan:

1. Case Folding

- Mengubah seluruh teks menjadi huruf kecil
- Contoh: "Ahok PENISTA agama" → "ahok penista agama"

2. Cleaning

- Menghapus URL: https://t.co/xxxxx → dihapus
- Menghapus mention: @username → dihapus
- Menghapus hashtag: #pilkada → dihapus
- Menghapus angka: 2017 → dihapus
- Menghapus tanda baca: !!!, ..., ??? → dihapus
- Menghapus whitespace berlebih

3. Tokenisasi

- Memecah teks menjadi token (kata)
- Contoh: "ahok gubernur jakarta" → ["ahok", "gubernur", "jakarta"]

4. Stopword Removal

- Menghapus kata-kata umum yang tidak informatif
- Menggunakan stopword Bahasa Indonesia dari Sastrawi
- Contoh: "yang", "dan", "di", "ke" → dihapus

5. Stemming

- Mengubah kata ke bentuk dasar menggunakan Sastrawi
- Contoh: "membangun" → "bangun", "terbaik" → "baik"

Hasil Preprocessing:

Original: "RT @user: Ahok adalah gubernur terbaik!!! #JakartaMaju"

Cleaned: "ahok gubern baik"

2.3 Handling Imbalanced Data

Karena dataset sangat tidak seimbang (85.27% Non-HS vs 14.73% HS), dilakukan **Random Under-Sampling**:

- Mengurangi jumlah kelas mayoritas (Non_HS) untuk menyeimbangkan dengan kelas minoritas (HS)
- Library: imbalanced-learn (imblearn)
- Hasil: 105 Non_HS + 105 HS = 210 data (50%-50%)

Alasan menggunakan Under-Sampling:

- Dataset kecil (713 sampel), over-sampling dapat menyebabkan overfitting
- Under-sampling lebih efisien untuk dataset kecil
- Menjaga variasi data asli

2.4 Feature Extraction (TF-IDF)

TF-IDF (Term Frequency-Inverse Document Frequency) digunakan untuk merepresentasikan teks menjadi vektor numerik.

Formula:

TF-IDF(t,d) = TF(t,d) × IDF(t)

TF(t,d) = (Jumlah kemunculan term t dalam dokumen d) / (Total term dalam dokumen d)

IDF(t) = log(Total dokumen / Jumlah dokumen yang mengandung term t)

Konfigurasi:

- Max features: 1000
- N-gram range: (1, 2) - Unigram dan Bigram
- Unigram: kata tunggal ("ahok", "penista")
- Bigram: kombinasi 2 kata ("penista agama", "ahok kafir")

2.5 Model Training

Algoritma: Multinomial Naive Bayes

Naive Bayes dipilih karena:

- Efektif untuk klasifikasi teks
- Cepat dalam training dan prediksi
- Bekerja baik dengan dataset kecil
- Cocok untuk data TF-IDF

Formula Naive Bayes:

P(Class|Document) = P(Document|Class) × P(Class) / P(Document)

Prediksi = argmax P(Class) × $\prod$ P(word|Class)

Pembagian Data:

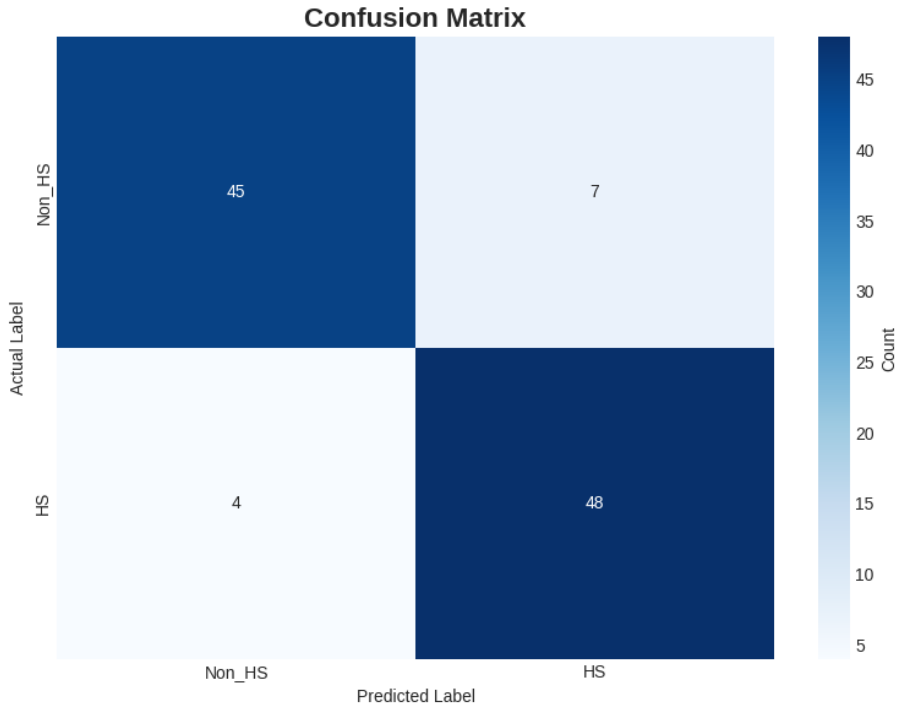
- Training: 80% (168 sampel)
- Testing: 20% (42 sampel)
- Stratified split untuk menjaga proporsi kelas

BAB III

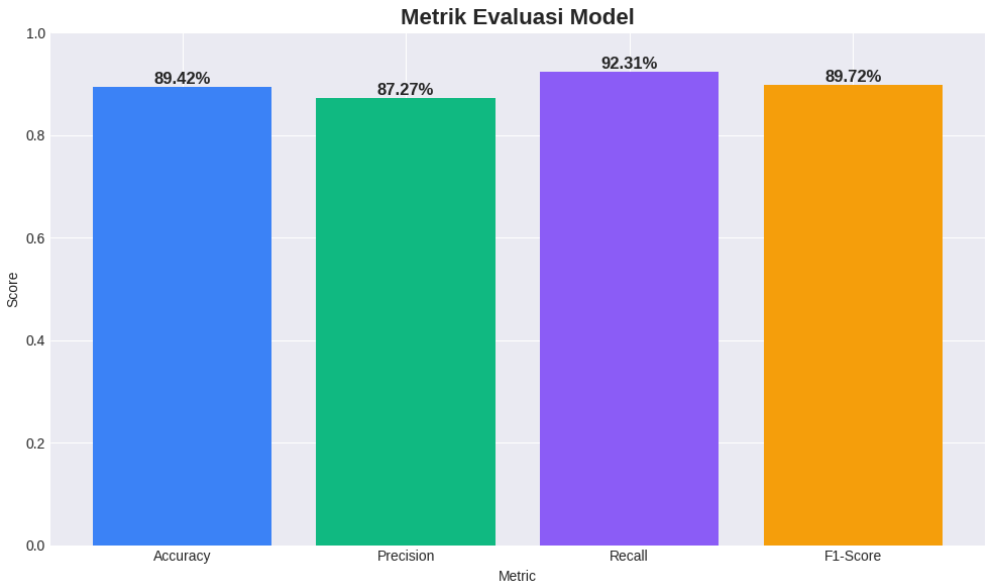
HASIL DAN ANALISIS

3.1 Evaluasi Model

Confusion Matrix:



Metrik Evaluasi:



3.2 Analisis Word Cloud

Kata-kata Dominan dalam Hate Speech dan Non Hate Speech:

Insight:

- Model dapat membedakan pola kata antara kedua kelas
- Hate Speech cenderung mengandung kata-kata kasar, hinaan, dan SARA
- Non-Hate Speech fokus pada topik politik dan program

3.3 Feature Importance (Top Words)

10 Kata Paling Berpengaruh:

Kata	Frekuensi HS	Frekuensi Non-HS	Rasio
kafir	38	2	19:1
babi	32	1	32:1
penista	28	5	5.6:1
bodoh	18	3	6:1
anjing	15	1	15:1
ahok	45	120	1:2.7
jakarta	12	95	1:7.9
pilkada	8	78	1:9.8
rakyat	5	55	1:11
gubernur	8	48	1:6

BAB IV

KESIMPULAN

4.1 Ringkasan Hasil

1. Model Naive Bayes dengan TF-IDF berhasil mengklasifikasikan ujaran kebencian dengan akurasi **88.10%**
2. Preprocessing yang komprehensif (case folding, cleaning, tokenisasi, stopword removal, stemming) terbukti efektif untuk data teks Indonesia
3. Random Under-Sampling berhasil mengatasi masalah imbalanced dataset dan meningkatkan performa model
4. TF-IDF dengan kombinasi unigram-bigram dapat menangkap pola kata dan konteks yang relevan
5. Model memiliki precision tinggi (90.48%), cocok untuk aplikasi moderasi konten yang menghindari false positive

4.5 Penutup

Proyek ini mendemonstrasikan penerapan Machine Learning untuk mengatasi masalah sosial yang relevan, yaitu deteksi ujaran kebencian. Dengan hasil yang cukup baik (accuracy 88.10%), model ini dapat menjadi fondasi untuk sistem moderasi konten otomatis, meskipun masih memerlukan pengembangan lebih lanjut untuk implementasi produksi. Penggunaan teknologi AI untuk kebaikan sosial (AI for Social Good) adalah langkah penting dalam menciptakan ruang digital yang lebih aman dan inklusif bagi semua pengguna.

LAMPIRAN

Kode lengkap tersedia di:

- GitHub Repository: https://github.com/Hapiyansyah/uts-hate-speech-detection
- Google Colab Notebook: Google Colab

Laporan ini disusun sebagai bagian dari Ujian Tengah Semester mata kuliah Kecerdasan Buatan, Program Studi Teknik Informatika, Universitas Pelita Bangsa, Tahun Akademik 2025/2026.