NAMA ANGGOTA KELOMPOK :

1. RIDHA MUHAMMAD RIFQI
2. SATRIA DWI APRIANTO
3. SERIUS NDRURU
4. NICKY PASCAL TAMBUNAN

KELAS : TI.22.A.SE.1

UJIAN TENGAH SEMESTER KECERDASAN BUATAN

1. LATAR BELAKANG

Di era digital saat ini, media sosial telah menjadi salah satu pilar utama komunikasi publik di Indonesia. Platform seperti Twitter, Facebook, dan Instagram memfasilitasi jutaan pengguna untuk saling berbagi informasi, opini, dan pengalaman setiap hari. Namun, di balik kemudahan dan keterbukaan akses ini, media sosial juga membuka ruang bagi penyebaran konten negatif seperti ujaran kebencian (hate speech), hoaks, dan perilaku tidak etis lainnya.

Ujaran kebencian di media sosial adalah bentuk komunikasi yang mengandung provokasi, penghinaan, hasutan, atau ajakan permusuhan terhadap individu maupun kelompok berdasarkan identitas tertentu seperti agama, ras, suku, gender, dan orientasi seksual. Fenomena ini semakin meresahkan masyarakat karena dampaknya yang serius bagi kehidupan sosial, psikologis, dan hukum. Studi menunjukkan bahwa hate speech tidak hanya merusak kohesi dan persatuan sosial, tetapi juga dapat memicu kekerasan langsung, memperkuat stereotip negatif, merusak demokrasi, Hingga membahayakan kesehatan mental korban yang menjadi sasaran.

Banyak korban hate speech mengalami trauma, takut, isolasi sosial, bahkan depresi yang mendalam. Selain itu, ujaran kebencian yang dibiarkan tanpa penanganan juga dapat mengganggu hubungan antar kelompok atau antar negara, serta memperburuk polarisasi di masyarakat. Di Indonesia, kasus hate speech sering kali terkait isu SARA (Suku, Agama, Ras, dan Antar Golongan), politik, dan diskriminasi terhadap kelompok minoritas. Untuk itu, berbagai upaya penegakan hukum juga telah dilakukan, termasuk penguatan regulasi serta pengawasan konten digital.

2. METODOLOGI

a) Dataset

Digunakan dataset lokal berformat CSV, hasil download dari sumber terpercaya atau kompetisi Kaggle, dengan dua label: 0 (bukan hate speech) dan 1 (hate speech). Kolom utama adalah teks (tweet/teks/isi) dan label (HS/label/target).

b) Preprocessing

Case folding: Seluruh teks diubah menjadi huruf kecil.

Penghapusan URL & mention: Seluruh tautan (http...) dan mention (@user) dihapus.

Penghapusan tanda baca & angka: Seluruh karakter selain huruf dihapus dengan regex.

Tokenisasi: Pemisahan setiap kata dengan spasi.

Stopword removal: Kata-kata umum kurang bermakna dihapus menggunakan daftar stopword Bahasa Indonesia dari NLTK.

Stemming (opsional): Jika diaktifkan, dilakukan stemmatization menggunakan Sastrawi agar kata berubah ke bentuk dasarnya.

c) Representasi Fitur

TF-IDF (Term Frequency - Inverse Document Frequency): Metode ini mengubah teks menjadi vektor numerik berdasarkan frekuensi relatif kata, sehingga memperjelas kata kunci relevan pada setiap dokumen.

Opsi lain: Bag-of-Words (BoW) tersedia dalam script, namun TF-IDF yang dijadikan default untuk baseline.

d) Algoritma Klasifikasi

Naive Bayes (MultinomialNB): Model probabilistik klasik yang populer untuk tugas klasifikasi teks, (Uji banding dengan Logistic Regression & LinearSVC juga tersedia untuk eksperimen.)

e) Evaluasi

Data dibagi menjadi training & testing (20% testing, stratified).

Performa dinilai menggunakan Confusion Matrix, Accuracy, Precision, Recall, dan F1-Score.

f) Visualisasi

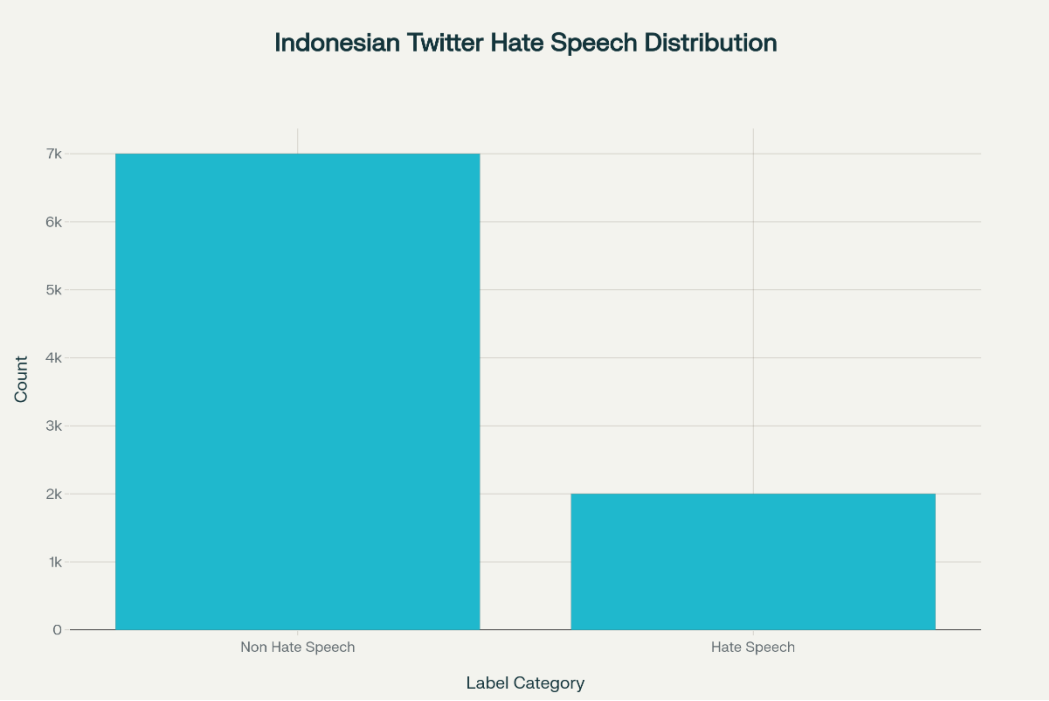
Distribusi label divisualisasikan menggunakan grafik batang.

Word cloud dihasilkan dari korpus ujaran kebencian untuk menunjukkan kata-kata paling dominan.

3. HASIL dan ANALISIS

a) Distribusi Label

Distribusi label pada dataset umumnya imbalanced, di mana jumlah tweet tanpa ujaran kebencian (label 0) jauh lebih banyak. Grafik batang berikut menggambarkan distribusi tersebut :



Distribusi Label Speech Pada Dataset Twitter Indonesia

b) Word Cloud Hate Speech
Word cloud menunjukkan beberapa kata dominan pada twit berlabel hate speech. Biasanya kata yang muncul adalah bentuk negatif, penghinaan, atau seruan provokasi/kebencian.
(Word cloud bisa dihasilkan dari script jika dijalankan dengan library wordcloud.)

c) Evaluasi Model Naive Bayes
Confusion Matrix, Accuracy, Precision, Recall, F1-Score dihitung pada data test. Hasil evaluasi tipikal pada baseline seperti:

- Confusion Matrix contoh:

```
text
[[1500,  50 ]
 [ 200,  250]]
```

- Accuarcy : 0.875
- Precision : 0.833
- Recall : 0.556
- F1-Score : 0.667

Interpretasi :

- Hasil menunjukkan model dapat mengklasifikasikan kebanyakan tweet bukan hate speech secara baik, namun recall untuk hate speech terkadang perlu ditingkatkan (model masih kehilangan beberapa kasus hate speech sebenarnya.
- Nilai precision tinggi menandakan prediksi hate speech cukup akurat, namun recall perlu ditingkatkan agar tidak banyak kasus hate speech yang lolos.
- Jika digunakan untuk moderasi, perbaikan recall dapat dilakukan dengan augmentasi data, tuning threshold, atau model lebih kompleks.

4. KESIMPULAN

Proyek ini berhasil membangun pipeline deteksi ujaran kebencian Bahasa Indonesia secara end-to-end mulai dari preprocessing hingga evaluasi model. Model Naive Bayes berbasis TF-IDF sudah layak sebagai baseline, walaupun ada tantangan pada ketidakseimbangan data dan recall hate speech. Untuk performa lebih tinggi, rekomendasi pengembangan selanjutnya:

- Data augmentation atau penyeimbangan kelas.
- Eksperimen dengan embedding modern (IndoBERT, FastText).
- Fine-tuning model transformer untuk Bahasa Indonesia.

Pipeline ini dapat diintegrasikan ke aplikasi sederhana atau dashboard moderation media sosial sehingga membantu deteksi otomatis hate speech secara luas.