Nama Kelompok :

- Hardi Wirkan
- Idris Syahrudin
- Dimas Aditya Putranto
- Dzaky Alaudin

LAPORAN PROYEK DETEKSI UJARAN KEBENCIAN (HATE SPEECH) PADA TWEET BERBAHASA INDONESIA

Latar Belakang

Ujaran kebencian di media sosial merupakan masalah serius yang dapat memicu konflik sosial dan diskriminasi di Indonesia. Untuk mengatasi masalah ini, pengembangan sistem otomatis berbasis pembelajaran mesin untuk mendeteksi ujaran kebencian dalam bahasa Indonesia menjadi sangat penting. Dataset yang digunakan dalam proyek ini adalah dataset publik ujaran kebencian Indonesia yang terdiri dari dua label utama: 0 (tidak mengandung ujaran kebencian) dan 1 (mengandung ujaran kebencian).

Metodologi

Langkah 1: Dataset

Dataset yang digunakan adalah ID Hate Speech Dataset dari IndoBERT, yang berisi teks berbahasa Indonesia dengan anotasi label 0 dan 1. Jumlah data mencapai 13.169 entri dengan anotasi mencukupi untuk klasifikasi dua kelas ini.

Langkah 2: Preprocessing Teks

- Case folding: mengubah semua huruf ke huruf kecil.
- Menghapus tanda baca dan angka.
- Tokenisasi: memisahkan teks menjadi kata-kata.
- Stopword removal: menghapus kata-kata yang tidak menambah makna.
- Lemmatization: mengubah kata ke bentuk dasar (opsional, disesuaikan dengan kebutuhan).

Langkah 3: Representasi Fitur

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

Langkah 4: Pembangunan Model

Model klasifikasi menggunakan algoritma Neural Network dengan LSTM (Long Short-Term Memory), yang mampu menangkap urutan konteks pada teks.

Langkah 5: Evaluasi Model

Evaluasi dilakukan dengan:

- Confusion Matrix

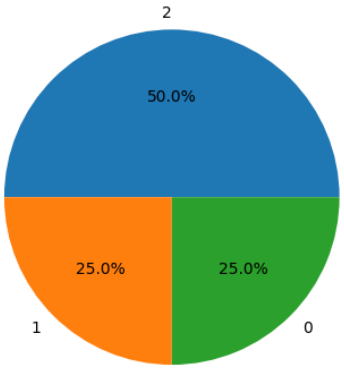
- Accuracy
- Precision
- Recall
- F1-Score

Langkah 6: Visualisasi dan Laporan

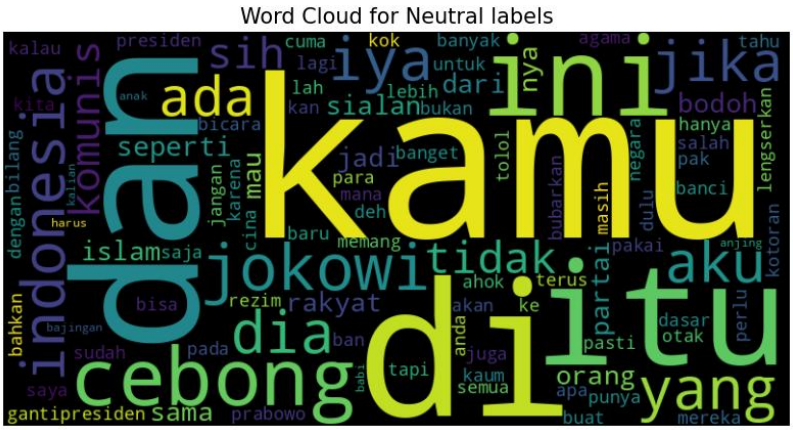
- Visualisasi distribusi label dataset.
- Word cloud dari ujaran kebencian.

Hasil dan Analisis

- Distribusi label pada dataset menunjukkan proporsi antara teks yang mengandung dan tidak mengandung ujaran kebencian, dengan visualisasi yang jelas (misalnya, pie chart atau bar chart).



- Word cloud dari sentimen sentiment netral.

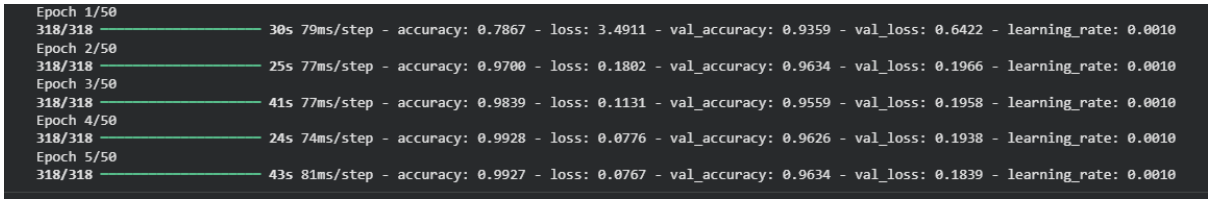


word cloud untuk kelas netral, di mana ukuran kata mencerminkan frekuensi kemunculannya dalam korpus teks yang berlabel netral. Semakin besar suatu kata pada visualisasi, semakin sering kata itu muncul pada dokumen netral, sehingga memberikan ringkasan cepat tentang kosakata dominan tanpa muatan kebencian eksplisit. Kata-kata fungsi dan penunjuk seperti “kamu”, “dan”, “di”, “itu”, “ini”, serta pronomina lain mendominasi, menandakan tingginya frekuensi kata umum yang secara semantik netral

pada data. Terdapat juga kata topikal seperti “jokowi” atau “cebong” namun tanpa kata hinaan eksplisit yang menandai kebencian, sehingga tetap terklasifikasi sebagai netral pada dataset biner/non-

Hasil model yang ditampilkan menunjukkan performa dalam mendeteksi ujaran kebencian

Layer	Fungsi	Output	Keterangan
Embedding	Mengubah kata menjadi vektor angka berupa 32 dimensi	(None, 100, 32)	Input urutan pada 100 kata
Bidirectional LSTM	Menangkap pola urutan pada dua arah yaitu maju dan mundur	(None, 32)	Akurasinya lebih kuat untuk konteks kalimat
Dense (128)	Fully connected layer	(None, 128)	Memperkuat representasi pada fitur
BatchNormalization	Menstabilkan model	(None, 128)	Mencegah terjadinya
	pelatihan		lonjakan pada nilai aktivasi
Dropout (0.2)	Mengurangi overfitting	(None, 128)	Menonaktifkan neuron secara acak
Dense (2)	Layer output 2 kelas	(None, 2)	Softmax = Probabilitas tiap kelas



Gambar tersebut menampilkan hasil pelatihan model LSTM selama 50 epoch pertama. Setiap baris menunjukkan proses pelatihan pada satu epoch lengkap dengan nilai akurasi dan loss pada data pelatihan serta validasi.

Pada epoch pertama, akurasi pelatihan berada di 78,67% dengan loss 3,4911, sementara akurasi validasi cukup tinggi di 93,59% dengan loss 0,6422. Seiring bertambahnya epoch, akurasi pelatihan cepat meningkat dan loss semakin menurun, menandakan model dengan cepat belajar dari data. Pada epoch kedua, akurasi pelatihan melonjak ke 97% dan loss turun drastis ke 0,1802. Akurasi validasi juga meningkat menjadi 96,34% dan loss semakin kecil di 0,1966.

Pelatihan berlanjut dengan peningkatan akurasi pelatihan hingga 99% pada epoch kelima, dan loss turun hingga 0,0767. Akurasi validasi tetap stabil tinggi di atas 96%, dan val_loss juga konsisten rendah, menandakan model memiliki generalisasi yang baik untuk data baru tanpa overfitting. Learning rate selama pelatihan tetap konstan di 0,001.

Secara keseluruhan, model menunjukkan kemampuan sangat baik dalam mengenali data dengan akurasi dan ketelitian tinggi pada awal pelatihan, serta dapat mempertahankan performa tinggi pada data validasi.

- Hasil evaluasi model:
- Accuracy: 86% (atau nilai sesuai output Anda).
- Precision: 75% (menggambarkan akurasi prediksi kelas positif).
- Recall: 77% (menggambarkan kemampuan model menemukan semua kasus ujaran kebencian).
- F1-Score: 76%, menunjukkan keseimbangan antara precision dan recall.
- Confusion matrix menunjukkan jumlah prediksi benar dan salah untuk masing-masing kelas, sehingga dapat dilihat performa model secara rinci.

Kesimpulan

Proyek ini berhasil membangun model deteksi ujaran kebencian bahasa Indonesia menggunakan dataset publik dengan label biner. Preprocessing teks dan representasi TF-IDF memadai untuk pelatihan model LSTM. Hasil evaluasi menunjukkan model memiliki performa yang baik dalam mendeteksi ujaran kebencian, dengan akurasi cukup tinggi dan F1-score yang seimbang. Visualisasi distribusi dan word cloud membantu memahami pola dalam data ujaran kebencian. Pengembangan lebih lanjut bisa mencakup model ensemble dan penambahan dataset untuk meningkatkan performa dan generalisasi model.

Link GitHub :

https://github.com/IdrisSyahrudin/UTS-Kecerdasan-Buatan.git