

Nama : Anindha Latiefa Zahra

NIM : 312210323

Kelas : TI.22.A.SE.1

UTS KECERDASAN BUATAN

LAPORAN MINI: Analisis Ujaran Kebencian Menggunakan Logistic Regression

1. Latar Belakang

Ujaran kebencian (hate speech) di media sosial merupakan salah satu permasalahan serius yang dapat menimbulkan dampak negatif terhadap individu maupun kelompok tertentu. Penyebaran ujaran kebencian tidak hanya mengancam keharmonisan sosial, tetapi juga dapat memicu konflik dan diskriminasi. Oleh karena itu, diperlukan sistem otomatis yang mampu mendeteksi ujaran kebencian secara cepat dan akurat.

Dalam proyek ini dilakukan implementasi sistem deteksi ujaran kebencian menggunakan machine learning, khususnya algoritma Logistic Regression, dengan memanfaatkan dataset publik “Hate Speech and Offensive Language Dataset”. Dataset tersebut berisi ribuan tweet yang telah diberi label berdasarkan kategori: ujaran kebencian, bahasa ofensif, atau netral. Tujuan utama dari penelitian ini adalah mengidentifikasi pola teks yang berpotensi mengandung ujaran kebencian dan mengevaluasi performa model klasifikasi yang digunakan.

2. Metodologi

2.1 Dataset

Dataset yang digunakan adalah `labeled_data.csv`, yang berisi lebih dari 4.900 data teks dari Twitter. Kolom utama yang digunakan yaitu:

tweet : isi teks tweet,

class : label kelas hasil anotasi dengan tiga kategori:

0 = Hate Speech (ujaran kebencian)

1 = Offensive Language (bahasa ofensif)

2 = Neither (tidak termasuk keduanya)

2.2 Preprocessing Teks

Agar data teks siap diproses oleh model machine learning, dilakukan beberapa tahap preprocessing:

1. Case Folding – mengubah semua huruf menjadi huruf kecil.
2. Menghapus angka dan tanda baca menggunakan modul ‘re’ dan ‘string’.
3. Tokenisasi dan Stopword Removal dengan menghapus kata umum seperti “the”, “and”, “you”.
4. Lemmatization untuk mengubah kata ke bentuk dasarnya, misalnya “running” → “run”.

5. Hasil teks bersih disimpan dalam kolom baru 'clean_text'.

2.3 Ekstraksi Fitur

Teks yang telah dibersihkan dikonversi menjadi vektor numerik menggunakan metode TF-IDF (Term Frequency – Inverse Document Frequency) dengan 5.000 fitur paling relevan. Metode ini menonjolkan kata yang sering muncul dalam dokumen tertentu namun jarang di seluruh korpus, sehingga lebih representatif.

2.4 Pembagian Data dan Model

Dataset dibagi menjadi:

- 80% data latih (training)
- 20% data uji (testing)

Model yang digunakan adalah Logistic Regression dengan 'max_iter=200'. Model ini dipilih karena sederhana, cepat, dan sering digunakan dalam klasifikasi teks.

2.5 Evaluasi Model

Model dievaluasi menggunakan beberapa metrik:

- Confusion Matrix
- Precision, Recall, F1-Score
- Accuracy

3. Hasil dan Analisis

3.1 Distribusi Label Dataset

Visualisasi pie chart menunjukkan bahwa dataset didominasi oleh kategori Offensive Language (kelas 1), mencapai lebih dari 75% dari total data. Sementara itu, kategori Hate Speech (kelas 0) dan Neither (kelas 2) memiliki proporsi lebih kecil.

Distribusi Kelas:

- Hate Speech (0): ± 6%
- Offensive Language (1): ± 77%
- Neither (2): ± 17%

Distribusi yang tidak seimbang ini berdampak pada performa model, karena model cenderung lebih akurat dalam mengenali kelas mayoritas.

3.2 Word Cloud Ujaran Kebencian

Dari word cloud yang dihasilkan untuk kelas Hate Speech, tampak bahwa kata-kata yang paling sering muncul adalah istilah bernada negatif, penghinaan, dan ekspresi kebencian terhadap kelompok tertentu. Visualisasi ini membantu memahami karakteristik leksikal dari ujaran kebencian, misalnya penggunaan kata-kata kasar atau diskriminatif yang menjadi indikator utama dalam deteksi otomatis.

3.3 Evaluasi Model

Berikut hasil pengujian model Logistic Regression:

Metrik	Nilai
Akurasi	0.896 (89.6%)
Presicion (kelas 1)	0.91
Recall (kelas 1)	0.96
F1-score (kelas 1)	0.94

Confusion matrix menunjukkan bahwa model sangat baik dalam mengenali kategori Offensive Language, namun masih lemah dalam mendekripsi Hate Speech karena jumlah datanya yang relatif sedikit. Nilai macro average F1 sebesar 0.69 menandakan performa cukup baik secara keseluruhan, meskipun masih ada ketidakseimbangan antar kelas.

4. Kesimpulan

Berdasarkan hasil eksperimen yang telah dilakukan, dapat disimpulkan bahwa:

1. Model Logistic Regression mampu mengklasifikasikan tweet dengan akurasi tinggi ($\pm 89,6\%$).
2. Distribusi label menunjukkan dominasi kategori offensive language, yang menyebabkan model lebih sensitif terhadap kelas tersebut.
3. Word cloud membantu mengidentifikasi kata-kata umum yang sering digunakan dalam ujaran kebencian, memberikan wawasan linguistik tambahan untuk analisis lebih lanjut.
4. Untuk meningkatkan akurasi deteksi hate speech, perlu dilakukan:
 - Penyeimbangan data (resampling atau data augmentation),
 - Eksperimen dengan model yang lebih kompleks seperti LSTM atau BERT,
 - Penyesuaian stopwords dan lemmatizer untuk bahasa spesifik (misalnya Bahasa Indonesia).

Dengan pendekatan ini, sistem deteksi ujaran kebencian dapat dikembangkan menjadi alat bantu yang lebih canggih dan relevan dalam memantau konten di media sosial.