Harap mengisi tabel ini, Tabel ini digunakan untuk keperluan komunikasi administrasi saja, saat publish akan dihapus oleh team editor.	
Nama Kontak	**Yoga Pratama**
Nomor WA	**0895338392281**
Prodi/Jurusan	**Teknik Informatika**
Perguruan Tinggi	**Universitas Pelita Bangsa**

KLASIFIKASI UJARAN KEBENCIAN PADA MEDIA SOSIAL MENGGUNAKAN TF-IDF DAN LOGISTIC REGRESSION

Yoga Pratama[1], Rhendy Diki Nugraha[2], Danang Nurcahyo[3], Ihsan Hadimulya[4]
Teknik Informatika, Universitas Pelita Bangsa
Jl. Inspeksi Kalimalang No.9, Cibatu, Bekasi, Indonesia
Yogafrtm2001@gmail.com

ABSTRAK

Perkembangan media sosial yang sangat pesat telah mempermudah masyarakat dalam berkomunikasi, tetapi juga memicu maraknya penyebaran ujaran kebencian (*hate speech*). Penelitian ini bertujuan untuk mengklasifikasikan ujaran kebencian pada media sosial menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF) dan algoritma *Logistic Regression*. Dataset yang digunakan terdiri dari dua kelas, yaitu *hate speech* dan *non–hate speech*. Tahapan penelitian meliputi *preprocessing* teks (pembersihan URL, angka, tanda baca, *case folding*, dan *stopword removal*), representasi fitur dengan TF-IDF, pelatihan model, serta evaluasi menggunakan metrik akurasi, presisi, *recall*, dan *F1-score*. Hasil penelitian menunjukkan bahwa model *Logistic Regression* mampu mendeteksi ujaran kebencian dengan performa yang baik dan seimbang di kedua kelas. Visualisasi berupa distribusi label memperlihatkan bahwa data relatif seimbang, sedangkan *word cloud* menampilkan kata-kata yang sering digunakan dalam ujaran kebencian seperti "kafir", "cebong", dan "anjing". Secara keseluruhan, kombinasi TF-IDF dan *Logistic Regression* terbukti efektif, sederhana, serta efisien untuk mendeteksi ujaran kebencian, dan dapat dijadikan dasar bagi pengembangan sistem moderasi konten otomatis pada media sosial.

Kata kunci : *Hate Speech, TF-IDF, Logistic Regression, Natural Language Processing, Klasifikasi Teks, Media Sosial*

1. PENDAHULUAN

Perkembangan Media sosial saat ini menjadi wadah utama bagi masyarakat untuk berkomunikasi, berbagi informasi, dan berpendapat. Namun, kebebasan tersebut juga memunculkan maraknya ujaran kebencian (*hate speech*) yang dapat mengganggu stabilitas sosial dan memicu konflik daring. Oleh karena itu, sistem klasifikasi otomatis diperlukan untuk membantu mendeteksi ujaran kebencian secara cepat dan akurat.

Berbagai penelitian telah dilakukan untuk mengatasi permasalahan ini. Logistic Regression telah terbukti efektif untuk klasifikasi sentimen pada Twitter dalam dataset terbatas [1]. Tantangan terbesar dalam deteksi ujaran kebencian adalah ketidakseimbangan data (*data imbalance*), yang sering diselesaikan dengan teknik seperti SMOTE [2]. Model berbasis Transformer seperti DistilBERT juga telah menunjukkan performa tinggi untuk klasifikasi hate speech [3], namun model tradisional masih menjadi pilihan karena kecepatan dan efisiensinya.

Studi di Indonesia menunjukkan bahwa Logistic Regression dapat digunakan untuk mendeteksi ujaran kebencian pada media sosial lokal dengan cukup baik [4]. Pendekatan lain seperti K-Nearest Neighbor yang dikombinasikan dengan TF-IDF juga memberikan hasil yang kompetitif [5]. Selain itu, beberapa penelitian telah menggunakan teknik ensemble dengan TF-IDF untuk meningkatkan akurasi deteksi [6]. Perbandingan antara BoW dan TF-IDF juga menunjukkan bahwa TF-IDF menghasilkan performa lebih baik pada tweet *real-time* [7].

Penelitian terbaru menegaskan bahwa metode tradisional dan ensemble dapat bersaing dengan model deep learning untuk klasifikasi ujaran kebencian [8]. Implementasi CNN untuk hate speech berbahasa Indonesia juga menunjukkan performa menjanjikan [9]. Studi lain membahas tantangan umum dalam deteksi hate speech Bahasa Indonesia, termasuk ambiguitas bahasa, sarkasme, dan konteks budaya [10].

Berdasarkan penelitian-penelitian tersebut, penggunaan TF-IDF dan Logistic Regression tetap menjadi pendekatan yang relevan karena hasil yang stabil dan interpretasi yang mudah. Oleh karena itu, penelitian ini mengimplementasikan kedua metode tersebut untuk mengklasifikasikan ujaran kebencian pada media sosial.

2. TINJAUAN PUSTAKA

2.1. Ujaran Kebencian

Ujaran kebencian merupakan ekspresi komunikasi yang menyerang individu atau kelompok berdasarkan ras, agama, etnis, atau karakteristik lain yang dapat memicu konflik sosial. Deteksi otomatis ujaran kebencian telah menjadi fokus utama berbagai penelitian karena meningkatnya aktivitas pengguna media sosial [10]. Tantangan utamanya adalah sifat bahasa yang kontekstual, penggunaan sarkasme, dan dinamika bahasa gaul yang berubah-ubah [9], sehingga diperlukan pendekatan yang efektif untuk mengidentifikasi pola kebencian secara konsisten.

2.2. Text Preprocessing

Preprocessing merupakan langkah penting dalam *natural language processing* (NLP) sebelum data dapat dianalisis lebih lanjut. Tahap ini biasanya meliputi *case folding*, pembersihan tanda baca, tokenisasi, dan penghapusan stopword untuk memperbaiki kualitas representasi teks [1]. Penelitian lain juga menegaskan bahwa normalisasi teks meningkatkan akurasi model karena mengurangi noise dan variasi kata yang tidak penting [7]. Dalam penelitian klasifikasi hate speech, preprocessing menjadi krusial mengingat bentuk bahasa informal yang banyak digunakan oleh pengguna media sosial [10].

2.3. TF-IDF

TF-IDF adalah metode pembobotan kata yang banyak digunakan untuk representasi fitur dalam klasifikasi teks. TF-IDF mampu menonjolkan kata-kata penting dan menurunkan bobot kata yang sering muncul tetapi tidak relevan, sehingga dapat meningkatkan performa model klasifikasi [5]. Perbandingan metode representasi teks menunjukkan bahwa TF-IDF lebih stabil dibandingkan Bag of Words dalam mendeteksi ujaran kebencian pada tweet real-time [7]. TF-IDF juga sering dipadukan dengan model tradisional maupun ensemble untuk meningkatkan performa [6].

2.4. Logistic Regression

Logistic Regression merupakan algoritma klasifikasi linear yang banyak digunakan dalam penelitian NLP karena kemampuannya menangani data berdimensi tinggi dan sparsity yang tinggi pada representasi teks [4]. Penelitian menunjukkan bahwa Logistic Regression mampu memberikan performa yang baik pada dataset terbatas dan seimbang [1]. Selain itu, berbagai studi juga membandingkan Logistic Regression dengan model deep learning, dan hasilnya menunjukkan bahwa metode tradisional ini tetap kompetitif untuk klasifikasi hate speech [8]. Implementasi Logistic Regression yang efisien menjadikannya pilihan tepat untuk sistem deteksi ujaran kebencian berskala besar [2].

3. METODE PENELITIAN

Metode penelitian yang digunakan dalam studi ini dimulai dengan pengumpulan dataset teks yang berisi dua label, yaitu ujaran kebencian dan non–ujaran kebencian. Data kemudian melalui proses *preprocessing* yang meliputi *case folding*, pembersihan URL, angka serta tanda baca, tokenisasi, dan penghapusan *stopword* Bahasa Indonesia. Teks yang telah dibersihkan selanjutnya diubah menjadi representasi numerik menggunakan metode TF-IDF. Setelah itu, model klasifikasi dibangun menggunakan algoritma Logistic Regression dengan komposisi pembagian data 80% untuk pelatihan dan 20% untuk pengujian. Evaluasi model dilakukan menggunakan metrik akurasi, presisi, *recall*, *F1-score*, serta *confusion matrix* untuk mengetahui performa prediksi tiap kelas. Visualisasi distribusi label dan *word cloud* juga digunakan untuk memahami karakteristik data serta pola kata yang sering muncul dalam ujaran kebencian.

4. HASIL DAN PEMBAHASAN

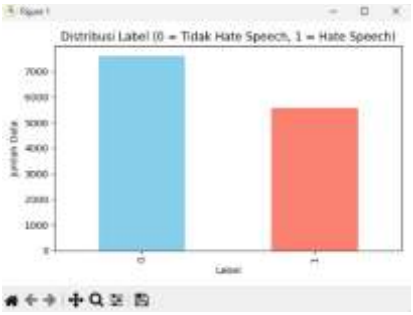
Penelitian ini menerapkan algoritma Logistic Regression dengan representasi fitur TF-IDF untuk klasifikasi ujaran kebencian pada media sosial. Data yang digunakan merupakan kumpulan teks berlabel dua kelas, yaitu *hate speech* (1) dan *non–hate speech* (0).

Pra-Pemrosesan dan Representasi Fitur

Teks pada dataset terlebih dahulu melalui tahap *preprocessing* yang mencakup *case folding*, penghapusan URL, angka, dan tanda baca, serta penghapusan *stopword* Bahasa Indonesia. Tahapan ini, sebagaimana diimplementasikan dalam fungsi clean_text, berhasil mengurangi noise dan membuat data lebih siap untuk dikonversi ke bentuk numerik menggunakan TF-IDF Vectorizer. Metode ini memberikan bobot lebih tinggi pada kata-kata yang penting dan menurunkan bobot kata umum, sehingga membantu model membedakan ujaran kebencian dari teks netral.

Distribusi Label Data

Sebelum pelatihan, dilakukan analisis distribusi label sebagaimana ditunjukkan pada Gambar 1. Jumlah data untuk kelas 0 (*tidak hate speech*) lebih banyak dibandingkan kelas 1 (*hate speech*), tetapi perbedaannya tidak ekstrem. Dengan demikian, pembagian data masih memungkinkan model belajar tanpa teknik penyeimbangan seperti SMOTE.



Gambar 1. Grafik batang jumlah data untuk label 0 (*tidak hate speech*) dan 1 (*hate speech)*

Analisis Word Cloud Ujaran Kebencian

Hasil Visualisasi *word cloud* pada Gambar 2 memperlihatkan kata-kata yang paling sering muncul

pada teks *hate speech*. Kata "user" muncul paling dominan, diikuti istilah bernada negatif seperti "kafir", "cebong", "anjing", "tolol", dan "cina". Pola ini menunjukkan bahwa ujaran kebencian di media sosial Indonesia kerap berkaitan dengan isu agama, politik, dan identitas sosial.

Gambar 2. Word cloud kata yang sering muncul pada kategori Hate Speech (Label = 1)

Kinerja Model

Model Logistic Regression yang dilatih dengan data TF-IDF menghasilkan nilai akurasi, presisi, recall, dan F1-score yang seimbang antara kedua kelas, yaitu *hate speech* dan *non–hate speech*. Hasil tersebut menunjukkan bahwa model tidak hanya fokus pada kelas mayoritas, tetapi juga memiliki kemampuan yang baik dalam mengenali ujaran kebencian yang jumlah datanya lebih sedikit. Nilai metrik evaluasi yang seimbang menandakan bahwa model mampu melakukan generalisasi dengan baik terhadap data uji yang belum pernah dilihat sebelumnya.

Confusion matrix yang dihasilkan memperlihatkan bahwa jumlah prediksi benar terhadap kedua kelas cukup tinggi, dengan tingkat kesalahan klasifikasi yang rendah. Hal ini mengindikasikan bahwa representasi teks menggunakan TF-IDF mampu menonjolkan kata-kata penting yang berpengaruh terhadap keputusan klasifikasi, seperti kata-kata bermakna kasar atau menyerang yang sering muncul pada teks *hate speech*. Keberhasilan model dalam membedakan pola bahasa kebencian dan non-kebencian juga menunjukkan bahwa kombinasi antara pendekatan *statistical feature extraction* dan model linier masih sangat relevan untuk kasus klasifikasi teks berbahasa Indonesia.

Secara keseluruhan, kombinasi TF-IDF dan Logistic Regression terbukti efektif, sederhana, serta efisien dalam mendeteksi ujaran kebencian di media sosial. Pendekatan ini mudah diimplementasikan, memiliki waktu komputasi cepat, dan memberikan hasil yang kompetitif dibandingkan metode *deep learning* yang lebih kompleks. Selain itu, model ini dapat dijadikan solusi awal atau *baseline* bagi pengembangan sistem moderasi konten otomatis di platform media sosial. Untuk penelitian selanjutnya, model dapat ditingkatkan dengan menambahkan teknik *feature selection*, *n-gram representation*, atau penerapan algoritma lain seperti SVM, LSTM, maupun BERT untuk menangkap konteks semantik yang lebih dalam dan meningkatkan performa deteksi ujaran kebencian.

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil menerapkan metode TF-IDF dan Logistic Regression untuk melakukan klasifikasi ujaran kebencian pada media sosial. Melalui tahapan *preprocessing* teks, representasi fitur menggunakan TF-IDF, serta pelatihan model dengan pembagian data secara *stratified*, model mampu mengenali pola bahasa yang membedakan antara teks ujaran kebencian dan *non–hate speech*. Berdasarkan hasil evaluasi, model menunjukkan performa yang baik dengan nilai akurasi, presisi, *recall*, dan *F1-score* yang seimbang pada kedua kelas. Visualisasi berupa distribusi label dan *word cloud* turut memberikan gambaran pola kata yang sering muncul dalam ujaran kebencian dan membantu memahami karakteristik data. Secara keseluruhan, kombinasi TF-IDF dan Logistic Regression terbukti menjadi pendekatan yang efektif, sederhana, cepat, serta efisien untuk mendeteksi ujaran kebencian di media sosial. Metode ini dapat diimplementasikan sebagai dasar pengembangan sistem moderasi konten otomatis maupun sebagai *baseline* untuk penelitian lanjutan dengan algoritma lain seperti SVM, LSTM, atau BERT yang mampu menangkap konteks semantik yang lebih dalam. Penelitian selanjutnya disarankan untuk memperluas variasi dataset serta menguji teknik *feature selection*, *n-gram*, dan *word embedding* guna meningkatkan performa klasifikasi secara signifikan.

DAFTAR PUSTAKA

Wajib menggunakan Mendeley style IEEE
Minimal jumlah referensi 10 buah, 5 tahun terakhir

[1] A. A. Putri *et al.*, "PENERAPAN METODE LOGISTIC REGRESSION UNTUK," vol. 7, no. 1, pp. 95–107.

[2] H. Sutanto and A. Puji, "Resolving Data Imbalance using SMOTE for the Analysis and Prediction of Hate Speech Sentences," vol. 02, pp. 198–203, 2025, doi: 10.14710/vol15iss2pp198-203.

[3] S. Saseendran, R. Sudharshan, V. Sreedhar, and S. Giri, "Classification of Hate Speech and Offensive Content using an approach based on DistilBERT".

[4] "No Title," pp. 1–13.

[5] N. A. Saputra, K. Aeni, and N. M. Saraswati, "Indonesian Hate Speech Text Classification Using Improved K-Nearest Neighbor with TF-IDF- ICSρF," vol. 11, no. 1, pp. 21–30, 2024, doi: 10.15294/sji.v11i1.48085.

[6] R. Sathishkumar, T. Karthikeyan, P. Praveen, and S. M. Shamsundar, "Ensemble Text Classification with TF-IDF Vectorization for Hate Speech Detection in Social Media," no. November 2023, 2024, doi:

10.1109/ICSCAN58655.2023.10395354.

[7] S. Akuma, T. Lubem, and I. T. Adom, "Comparing Bag of Words and TF - IDF with different models for hate speech detection from live tweets," *Int. J. Inf. Technol.*, no. September, 2022, doi: 10.1007/s41870-022-01096-4.

[8] R. L. Hasanah *et al.*, "Perbandingan Tradisional dan Ensemble Machine Learning dalam Melakukan Klasifikasi Kalimat Ujaran Kebencian," vol. 08, no. 02, pp. 121–131, 2023.

[9] D. Ayu, N. Taradhita, I. K. Gede, and D. Putra, "Hate Speech Classification in Indonesian Language Tweets by Using Convolutional Neural Network," vol. 14, no. 3, pp. 225–239, 2021, doi: 10.5614/itbj.ict.res.appl.2021.14.3.2.

[10] E. W. Pamungkas, D. Galih, P. Putri, and A. Fatmawati, "Hate Speech Detection in Bahasa Indonesia : Challenges and Opportunities," no. August, 2023, doi: 10.14569/IJACSA.2023.01406125.