DETEKSI HATE SPEECH BERBASIS MACHINE LEARNING MENGGUNAKAN LINEAR SUPPORT VECTOR MACHINE

Johanes Mula Febrian Sihombing[1], Muhamad Abdul Anas[2], Gilang Rizki Bahtiar[3]
[1, 2, 3]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa
Email_Penulis : johanesmula@mhs.pelitabangsa.ac.id

ABSTRAK

Penelitian ini membahas deteksi ujaran kebencian pada media sosial menggunakan model *Linear Support Vector Machine*. Masalah ujaran kebencian perlu diperhatikan karena dapat memicu konflik dan mempengaruhi opini publik. Tujuan penelitian ini yaitu mengklasifikasikan teks ke dalam dua kategori yaitu normal dan *hate speech*. Dataset yang digunakan terdiri dari dua label dengan jumlah data normal sebanyak 7608 dan *hate speech* sebanyak 5561 data. Metode yang digunakan yaitu ekstraksi fitur TF IDF dan pemodelan menggunakan Linear SVM. Data dilatih dengan pembagian data *train* dan *test* untuk melihat kinerja model dalam mendeteksi pola bahasa pada dua kelas. Hasil pengujian menunjukkan bahwa model menghasilkan nilai akurasi sebesar 0.84. Nilai precision dan recall pada dua kelas juga relatif seimbang sehingga menunjukkan kemampuan model dalam mempelajari pola kata pada setiap kategori. Hasil ini memperlihatkan bahwa kombinasi TF IDF dan Linear SVM dapat digunakan sebagai pendekatan dasar untuk proses identifikasi ujaran kebencian terutama pada data berbahasa Indonesia. Pendekatan ini diharapkan dapat mendukung pembuatan sistem moderasi otomatis pada platform digital untuk membantu proses filtrasi konten berbasis teks.

Kata kunci : *ujaran kebencian, TF IDF, klasifikasi teks, Linear SVM, machine learning.*

1. PENDAHULUAN

Ujaran kebencian menyebar melalui media sosial dalam volume sangat besar setiap hari. Komentar menyerang kelompok identitas tertentu dapat memicu provokasi, diskriminasi dan konflik antar pengguna. Platform digital terus berkembang dan jumlah pengguna aktif terus naik. Kondisi ini membuat proses moderasi manual tidak mampu mengimbangi kecepatan arus konten baru. Moderator manusia akan selalu tertinggal dari banjir data, karena konten baru masuk setiap detik. Kamu membutuhkan sistem otomatis yang mampu menandai konten berbahaya sejak awal sehingga keputusan moderasi dapat diambil lebih cepat.Nama pertama nama mahasiswa, nama kedua dan ketiga adalah dosen pembimbing pertama dan kedua tanpa gelar, alamat email cukup satu saja gunakan email mahasiswa atau Corresponding Author [1].

Analisis manual juga memiliki tantangan lain. Bahasa yang digunakan pengguna media sosial tidak formal. Ada singkatan, campuran bahasa, simbol, typo dan gaya bahasa ironi. Makna kebencian sering muncul dalam bentuk kalimat pendek. Ada kalimat yang terlihat biasa secara permukaan, tetapi memiliki konteks serangan ketika berada dalam topik tertentu. Perbedaan gaya penulisan membuat manusia harus membaca satu per satu posting untuk menentukan apakah suatu konten mengandung kebencian atau tidak. Pendekatan manual membutuhkan waktu dan biaya besar [2].

Machine learning dapat memberikan solusi. Model dapat belajar dari dataset yang sudah memiliki label. Setiap kalimat dapat dikonversi menjadi fitur numerik lalu dianalisis untuk menemukan pola kata dan pola kombinasi kata yang sering muncul pada teks yang mengandung ujaran kebencian. Model kemudian dapat memprediksi kategori teks baru tanpa intervensi manusia. Pendekatan ini memberikan proses moderasi yang lebih cepat dan lebih konsisten. Selain itu model dapat dievaluasi dan diukur secara kuantitatif menggunakan metrik evaluasi yang jelas.

Penelitian ini menggunakan dataset dua kelas. Kelas pertama adalah teks normal. Kelas kedua adalah teks yang mengandung ujaran kebencian. Teks dibersihkan terlebih dahulu melalui beberapa tahap preprocessing. Setelah itu teks diubah menjadi representasi numerik menggunakan TF IDF dengan batas 10000 fitur. Model Linear Support Vector Machine kemudian digunakan untuk memisahkan dua kelas tersebut. Model dievaluasi menggunakan accuracy, precision, recall, F1 Score dan confusion matrix. Pendekatan ini memberi gambaran jelas tentang seberapa baik model dalam membedakan teks normal dan teks yang mengandung hate speech pada dataset bahasa Indonesia.

2. TINJAUAN PUSTAKA

2.1. Hate Speech Detection

Ujaran kebencian adalah ekspresi negatif berisi hinaan atau serangan yang ditujukan kepada individu atau kelompok. Pemetaan ujaran kebencian pada platform digital dapat membantu moderasi konten. Pendekatan otomatis menggunakan machine learning digunakan untuk mengklasifikasikan teks ke dalam kategori hate speech atau normal. Model belajar dari contoh data yang sudah memiliki label sehingga model dapat mengenali pola kata yang muncul pada teks hate speech [3].

Akurasi deteksi sangat bergantung pada kualitas data yang digunakan. Semakin baik proses pelabelan

dataset maka semakin kuat pola yang dapat dipelajari model. Penelitian terdahulu banyak menggunakan pendekatan supervised learning untuk tugas ini karena model dapat mempelajari pola kata yang berkaitan dengan konteks ujaran kebencian. Proses ini membantu model untuk menandai posting atau komentar yang mengandung serangan secara otomatis. Implementasi sistem ini dapat membantu mempercepat proses moderasi pada platform media sosial yang memiliki volume data tinggi [4].

2.2. TF IDF

Representasi teks diperlukan agar data teks dapat diproses oleh algoritma machine learning. TF IDF mengubah teks menjadi fitur numerik. TF atau term frequency menghitung seberapa sering kata muncul pada dokumen. IDF atau inverse document frequency memberi bobot lebih besar pada kata yang jarang muncul pada seluruh dokumen. Fitur TF IDF dapat menonjolkan kata yang memiliki nilai pembeda antar kelas sehingga model dapat mempelajari hubungan antara kata dan label dengan lebih efektif.

TF IDF mengubah teks menjadi angka yang dapat diproses oleh algoritma machine learning. TF mengukur seberapa sering kata muncul dalam satu dokumen. IDF memberi bobot yang lebih besar untuk kata yang jarang muncul di keseluruhan dokumen. Perpaduan TF dan IDF menghasilkan bobot yang menonjolkan kata yang memiliki nilai pembeda antar dokumen. Pendekatan ini umum digunakan untuk representasi teks karena hasilnya stabil saat digunakan pada dataset berbahasa Indonesia.

Penggunaan TF IDF membantu model dalam memprioritaskan kata yang relevan saat melakukan klasifikasi. Kata dengan bobot tinggi memberikan informasi lebih kuat untuk membedakan teks normal dengan teks hate speech. Pendekatan ini sering digunakan pada penelitian analisis sentimen dan klasifikasi teks karena implementasinya sederhana dan mampu memberikan fitur yang cukup kuat untuk model klasifikasi [5].

2.3. Linear Support Vector Machine

Support Vector Machine adalah model klasifikasi yang mencari garis pemisah terbaik antara dua kelas. Linear SVM digunakan untuk kasus binary classification karena model ini mencari hyperplane linear yang dapat memaksimalkan margin pemisah antar kelas. Pendekatan linear menghasilkan proses training yang cepat pada data berukuran besar. Linear SVM menjadi salah satu algoritma yang efektif untuk klasifikasi teks karena ruang fitur teks cenderung berdimensi tinggi.

Linear Support Vector Machine mencari garis pemisah terbaik antara dua kelas pada ruang fitur. Model ini fokus mencari hyperplane yang menghasilkan margin maksimal antara data kelas berbeda. Semakin besar margin, semakin baik kemampuan model dalam memisahkan data. Linear SVM cocok digunakan untuk kasus binary classification. Linear SVM memiliki proses training yang relatif lebih cepat sehingga cocok untuk dataset besar.

Penggunaan Linear SVM pada teks efektif karena representasi teks memiliki dimensi fitur yang besar. Model linear dapat memanfaatkan bobot fitur untuk menentukan batas keputusan yang jelas antara kelas normal dan kelas hate speech. Linear SVM juga lebih mudah dievaluasi dan tidak memerlukan proses training yang kompleks seperti neural network. Model dapat bekerja baik pada dataset bahasa Indonesia apabila fitur yang digunakan mampu menangkap kata yang relevan dengan konteks hate speech [6].

3. METODE PENELITIAN

Penelitian ini menggunakan dataset dua kelas yang berisi teks normal dan teks yang mengandung ujaran kebencian. Dataset memiliki dua kolom utama yaitu Tweet sebagai teks mentah dan binary_label sebagai label kategori. Data ini berasal dari platform media sosial sehingga teks yang masuk tidak terstruktur. Variasi penggunaan bahasa informal membuat tahap pembersihan data menjadi penting. Dataset diperoleh dari platform Kaggle. Penelitian ini hanya fokus pada data bahasa Indonesia untuk menjaga konsistensi karakteristik bahasa.

Proses penelitian dimulai dengan pembersihan teks untuk menghilangkan elemen yang tidak relevan. Pembersihan meliputi penghapusan URL, mention, tanda baca dan karakter non huruf. Teks kemudian diubah menjadi huruf kecil dan stopword dihapus agar model fokus pada kata yang penting. Proses ini bertujuan untuk mengurangi noise pada data. Tokenisasi dilakukan untuk memisahkan kata menjadi unit yang lebih kecil sehingga memudahkan pemetaan kata ke fitur numerik. Tahap preprocessing dilakukan sebelum proses ekstraksi fitur.

Representasi fitur menggunakan TF IDF. Fitur ini mengubah teks menjadi angka sehingga dapat diproses oleh model machine learning. Jumlah fitur dibatasi 10000 fitur agar proses training lebih stabil dan tidak terlalu berat. Pendekatan ini dapat menonjolkan kata yang memiliki nilai pembeda antar kelas. TF IDF juga mengurangi pengaruh kata yang terlalu sering muncul tetapi tidak memiliki nilai informasi. Representasi ini dipilih karena hasilnya stabil pada dataset berbahasa Indonesia.

Model yang digunakan adalah Linear Support Vector Machine. Data dibagi menjadi data training dan data testing dengan rasio 80 persen data training dan 20 persen data testing. Model dilatih menggunakan data training lalu dievaluasi menggunakan data testing. Penelitian ini menggunakan pendekatan kuantitatif karena seluruh pengujian model dievaluasi menggunakan metrik numerik. Evaluasi menggunakan confusion matrix, accuracy, precision, recall dan F1 Score untuk melihat kemampuan model dalam mengklasifikasikan teks ke dalam dua kategori tersebut. Evaluasi berbasis metrik kuantitatif memudahkan peneliti dalam menilai kemampuan model saat memprediksi data baru dan melihat konsistensi model pada dua kelas yang berbeda.

4. HASIL DAN PEMBAHASAN

Dataset yang digunakan memiliki dua kelas yaitu normal dan hate speech. Kelas normal berjumlah 7608 data dan kelas hate speech berjumlah 5561 data. Jumlah ini menunjukkan bahwa data kelas normal lebih banyak dibandingkan data hate speech sehingga model perlu belajar memisahkan pola kata yang berbeda pada dua kelompok data ini. Jumlah data yang besar pada kelas normal berpengaruh pada komposisi fitur kata yang paling sering muncul pada proses ekstraksi TF IDF karena nilai frekuensi kata pada kelas normal akan lebih dominan pada proses pelatihan. Kondisi ini dapat menyebabkan ketidakseimbangan data yang berdampak pada proses pembelajaran model pada tahap training, sebab model akan lebih sering melihat pola kata dari kelas normal daripada kelas hate speech. Kondisi ketidakseimbangan ini perlu diperhatikan karena dapat menurunkan kemampuan model dalam mengenali pola teks hate speech yang jumlah datanya lebih sedikit [7].

Grafik distribusi label digunakan untuk melihat komposisi data pada masing masing kelas. Grafik memperlihatkan bahwa kelas normal mendominasi dibandingkan kelas hate speech. Ketidakseimbangan ini memberikan pengaruh pada model karena model lebih sering belajar dari pola kata yang berasal dari data normal. Kondisi ini dapat mengurangi kinerja model dalam mendeteksi kelas hate speech karena jumlah contoh kalimat hate speech lebih sedikit sehingga variasi kata negatif yang mewakili kategori hate speech menjadi lebih sedikit juga. Pada banyak penelitian text classification, ketidakseimbangan data seperti ini dapat menurunkan kemampuan model dalam mengenali pola kalimat yang mengandung ujaran kebencian, terutama jika bentuk ujaran kebencian ditulis dalam variasi kalimat yang berbeda. Distribusi data yang tidak seimbang juga dapat menyebabkan bias terhadap kelas mayoritas, karena model akan lebih mudah mengklasifikasikan teks baru sebagai kelas normal [8].

Gambar 1. Word cloud label 1

Wordcloud digunakan untuk menampilkan kata yang sering muncul pada kelas hate speech. Wordcloud hanya mengambil semua teks dari label 1 lalu menampilkan kata yang muncul paling banyak. Kata dengan ukuran lebih besar adalah kata yang memiliki frekuensi tinggi pada kelas hate speech. Wordcloud menunjukkan representasi visual dari pola kata yang sering muncul dalam ujaran kebencian, sehingga peneliti dapat melihat kata apa saja yang sering digunakan dalam konteks negatif. Pada hasil wordcloud terlihat beberapa kata yang sering muncul seperti kata “user” karena pengambilan data dari platform media sosial yang banyak menampilkan mention atau username. Selain itu muncul beberapa kata yang berkaitan dengan konteks politik di Indonesia seperti “cina”, “cebong”, “jokowi”, “prabowo”, dan beberapa kata penghinaan lain yang menunjukkan konteks ujaran kebencian politik. Temuan ini menunjukkan bahwa konten hate speech pada dataset ini dominan mengarah kepada isu politik lokal di Indonesia dan sering kali ditulis menggunakan bahasa informal dan campuran.

Tabel 1. Hasil evaluasi model Linear SVM

Metrik	Hasil
Accuracy	0.84
Precision normal	0.87
Precision hate speech	0.81
Recall normal	0.88
Recall hate speech	0.79
F1 normal	0.87
F1 hate speech	0.81

Hasil evaluasi model Linear SVM ditampilkan pada Tabel 1. Model menghasilkan nilai accuracy sebesar 0.84. Model mampu membedakan teks normal dan teks hate speech dengan cukup baik karena metode SVM bekerja dengan cara memaksimalkan margin antara dua kelas sehingga batas pemisah antar kelas menjadi jelas. Pendekatan linear yang digunakan oleh SVM dapat bekerja efektif pada representasi fitur yang tinggi seperti TF IDF karena metode ini sangat cocok untuk menangani data teks yang memiliki dimensi fitur sangat banyak namun sparsity tinggi. Hasil evaluasi menunjukkan bahwa model sudah mampu melakukan generalisasi terhadap data yang belum pernah dilihat sebelumnya karena model dapat mengenali pola kata pada dua kelas dengan performa yang stabil. Nilai akurasi ini menunjukkan bahwa model dapat digunakan untuk melakukan klasifikasi otomatis teks hate speech pada dataset yang memiliki karakteristik sejenis.

Confusion matrix menunjukkan 1338 data normal terdeteksi benar dan 881 data hate speech terdeteksi benar. Kesalahan terjadi pada 184 data normal dan 231 data hate speech. Hal ini menunjukkan bahwa sebagian teks hate speech masih memiliki pola kata yang mirip dengan kalimat normal sehingga model salah memprediksi. Perbedaan jumlah kesalahan prediksi pada kedua kelas menunjukkan bahwa model lebih mudah mengenali pola kalimat normal dibandingkan pola kalimat hate speech. Kondisi ini juga menunjukkan bahwa karakter bahasa hate speech yang digunakan dalam dataset cenderung bervariasi dan tidak selalu menggunakan kata kata kasar secara eksplisit. Selain itu banyak kalimat hate speech yang menggunakan konteks sarkasme, plesetan, singkatan, dan campuran bahasa sehingga model linear lebih sulit menemukan pola kata spesifik. Faktor faktor ini dapat menjadi penyebab mengapa

masih terjadi salah klasifikasi terutama pada label 1. Model memerlukan lebih banyak variasi data untuk menangkap pola bahasa yang lebih kompleks pada kelas hate speech [9].

Model Linear SVM sudah mampu memberikan performa yang cukup baik untuk klasifikasi dua kelas. Representasi fitur TF IDF memberikan kontribusi kuat dalam menonjolkan kata penting pada data dan membantu model melihat perbedaan bobot antar kata yang sering muncul pada kelas normal maupun hate speech. Metrik F1 pada dua kelas sudah seimbang sehingga model stabil dalam mengenali dua kategori dan tidak hanya unggul pada satu kelas saja. Hasil ini menunjukkan bahwa pendekatan ini dapat digunakan sebagai dasar pengembangan sistem moderasi otomatis pada platform media sosial dengan tingkat akurasi yang dapat diterima. Model dapat digunakan sebagai model dasar untuk pengembangan sistem yang mendukung deteksi ujaran kebencian pada data berbahasa Indonesia karena model mampu melakukan generalisasi dengan baik pada data uji. Pengembangan selanjutnya dapat dilakukan dengan melakukan optimasi fitur dan perbandingan metode model agar performa dapat ditingkatkan untuk penggunaan pada lingkungan yang lebih luas.

5. KESIMPULAN DAN SARAN

Penelitian ini menunjukkan bahwa model Linear Support Vector Machine mampu mendeteksi ujaran kebencian pada dataset dua kelas dengan nilai accuracy sebesar 0.84. Nilai precision, recall dan F1 pada dua kelas berada pada nilai yang relatif seimbang. Hasil ini memperlihatkan bahwa Linear SVM mampu mempelajari pola kata pada kalimat normal maupun kalimat hate speech secara cukup baik. Penggunaan TF IDF dengan batas 10000 fitur menjadi representasi yang efektif untuk memetakan teks menjadi vektor numerik. Evaluasi menggunakan confusion matrix memperlihatkan bahwa model mampu mengidentifikasi banyak data pada dua kelas dengan tepat, namun masih terdapat kesalahan prediksi karena terdapat kalimat yang memiliki pola kata mirip antara dua kelas. Secara umum, hasil yang diperoleh sudah menunjukkan bahwa model Linear SVM dapat digunakan sebagai baseline awal untuk deteksi ujaran kebencian berbasis teks berbahasa Indonesia.

Penelitian selanjutnya dapat meningkatkan kinerja model melalui beberapa langkah pengembangan. Peneliti dapat melakukan balancing data untuk mengatasi perbedaan jumlah data antar kelas agar model tidak cenderung memprediksi kelas normal. Peneliti juga dapat melakukan perbandingan performa antara Linear SVM dengan beberapa model lain seperti Logistic Regression dan Random Forest untuk melihat apakah terdapat peningkatan performa. Eksperimen hyperparameter tuning juga dapat dilakukan untuk melihat pengaruh parameter terhadap nilai akurasi. Selain itu dapat dilakukan penerapan teknik penggabungan fitur berbasis kamus kata atau lexicon untuk memfilter kata berbahaya yang sering muncul pada teks. Pengembangan sistem deteksi hate speech dapat diarahkan pada penerapan model ini ke media sosial secara real time agar mampu membantu proses moderasi konten secara otomatis dan lebih efisien.

DAFTAR PUSTAKA

[1] M. R. Anggana, “PERAN NATURAL LANGUAGE PROCESSING (NLP) DALAM MENGIDENTIFIKASI DAN MENGATASI BIAS GENDER PADA UJARAN KEBENCIAN,” *Jurnal Humaniora dan Teknologi*, vol. 11, no. 1, 2025.

[2] R. Stoleriu, A. Nascu, and F. Pop, “CYBERBULLYING DETECTION ON TIKTOK USING A DEEP LEARNING APPROACH,” *U.P.B. Sci. Bull., Series C*, vol. 87, no. 1, p. 2025, 2025.

[3] J. Khatib Sulaiman, A. Ariska, M. Kamayani, and U. Muhammadiyah DrHamka, “Deteksi Hate Speech pada Kolom Komentar Tiktok dengan menggunakan SVM,” *Indonesian Journal of Computer Science*, 2022.

[4] O. H. Rahman, G. Abdillah, and A. Komarudin, “Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine,” *Jurnal RESTI*, vol. 5, no. 1, pp. 17–23, Feb. 2021, doi: 10.29207/resti.v5i1.2700.

[5] O. I. Gifari, M. Adha, I. Rifky Hendrawan, F. Freddy, and S. Durrand, “Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine,” *JIFOTECH (JOURNAL OF INFORMATION TECHNOLOGY*, vol. 2, no. 1, 2022.

[6] N. Azmi Verdikha, R. Habid, and A. Johar Latipah, “Analisis DistilBERT dengan Support Vector Machine (SVM) untuk Klasifikasi Ujaran Kebencian pada Sosial Media Twitter,” *METIK JURNAL*, vol. 7, no. 2, pp. 101–110, Dec. 2023, doi: 10.47002/metik.v7i2.583.

[7] I. And and D. Expert, “Deteksi Ujaran Kebencian dengan Metode Klasifikasi Naïve Bayes dan Metode N-Gram pada Dataset Multi-Label Twitter Berbahasa Indonesia INFORMASI ARTIKEL ABSTRAK,” 2022. [Online]. Available: http://index.unper.ac.id

[8] N. M. Andini, Y. Findawati, I. R. I. Astutik, and A. Eviyanti, “Implementasi Convolutional Neural Network (CNN) Untuk Mendeteksi Ujaran Kebencian Dan Emosi Di Twitter,” *SMATIKA JURNAL*, vol. 14, no. 02, pp. 314–325, Dec. 2024, doi: 10.32664/smatika.v14i02.1346.

[9] D. Lazuardi, S. Putra, and T. Mauritsus, "SENTIMEN ANALYSIS ON THE NEW VARIANT OF COVID-19 (OMICRON) IN INDONESIA USING BERT TEXT REPRESENTATION," *J Theor Appl Inf Technol*, vol. 31, no. 14, 2023, [Online]. Available: www.jatit.org