HATE SPEECH CLASSIFICATION IN INDONESIAN SOCIAL MEDIA COMMENTS USING THE DECISION TREE ALGORITHM

Research · June 2025

DOI: 10.13140/RG.2.2.35551.57762

2 authors:

Muhammad Ghianza Al Ghifari
Universitas Sebelas April
4 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Fathoni Mahardika
Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Sumedang
53 PUBLICATIONS **91** CITATIONS
SEE PROFILE

Jurnal Teknik Informatika (JUTIF)
Vol. x, No. y, December 2023, pp. x-y

DOI: 10.52436/jutif...
p-ISSN: 2723-3863
e-ISSN: 2723-3871

HATE SPEECH CLASSIFICATION IN INDONESIAN SOCIAL MEDIA COMMENTS USING THE DECISION TREE ALGORITHM

Muhammad Ghianza Al Ghifari [*1], Fathoni Mahardika[2]

[1,2]Informatika, Fakultas Teknologi Informasi, Universitas Sebelas April, Indonesia
Email: [1]220660121058@student.unsap.ac.id, [2]fathoni@unsap.ac.id

(Article received: date; Revision: date; published: date)

Abstract

The growth of social media in Indonesia has triggered a rise in the spread of hate speech, negatively impacting social cohesion. This study aims to build and evaluate a hate speech classification model in Indonesian-language social media comments using the Decision Tree algorithm. The method involves data preprocessing, feature extraction using Term Frequency–Inverse Document Frequency (TF-IDF), model training, and performance evaluation. The dataset consists of 2,000 comments, with 60% labeled as hate speech and 40% as neutral. The model was trained with an 80:20 train-test split. Evaluation results show that the model achieved an accuracy of 89.50%, precision of 91.20%, recall of 87.60%, and F1-score of 89.35%. The model demonstrates competitive performance while offering good interpretability, making it suitable for automated content moderation systems. Nevertheless, challenges remain in detecting implicit or disguised hate speech. This research concludes that the Decision Tree algorithm can serve as an effective preliminary solution for detecting hate speech in Indonesian social media platforms.

Keywords: *Decision Tree, Hate speech, Indonesian Language, NLP, Social Media, Text Classification*

KLASIFIKASI UJARAN KEBENCIAN DALAM KOMENTAR MEDIA SOSIAL INDONESIA MENGGUNAKAN ALGORITMA DECISION TREE

Abstrak

Pertumbuhan media sosial di Indonesia memicu meningkatnya penyebaran ujaran kebencian yang berdampak negatif terhadap kohesi sosial. Penelitian ini bertujuan untuk membangun dan mengevaluasi model klasifikasi ujaran kebencian dalam komentar media sosial berbahasa Indonesia menggunakan algoritma *Decision Tree*. Metode yang digunakan mencakup tahapan *preprocessing* data, ekstraksi fitur menggunakan *Term Frequency–Inverse Document Frequency (TF-IDF)*, pelatihan model, dan evaluasi performa. *Dataset* terdiri dari 2.000 komentar dengan proporsi 60% ujaran kebencian dan 40% komentar netral. Model dilatih dengan rasio data 80:20 untuk pelatihan dan pengujian. Hasil evaluasi menunjukkan bahwa model mampu mencapai akurasi sebesar 89,50%, *precision* 91,20%, *recall* 87,60%, dan *F1-score* 89,35%. Model ini menunjukkan performa yang kompetitif sekaligus memberikan interpretabilitas yang baik, sehingga cocok untuk digunakan dalam sistem moderasi konten otomatis. Kendati demikian, tantangan masih ditemukan dalam mendeteksi ujaran kebencian implisit atau terselubung. Penelitian ini menyimpulkan bahwa algoritma *Decision Tree* dapat menjadi solusi awal yang efektif dalam mendeteksi ujaran kebencian di media sosial berbahasa Indonesia.

Kata kunci: *Bahasa Indonesia, Decision Tree, Klasifikasi Teks, Media Sosial, NLP, Ujaran Kebencian*

1. PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi telah mendorong pertumbuhan pesat penggunaan media sosial di Indonesia. Platform seperti Twitter, Facebook, dan Instagram tidak hanya menjadi sarana komunikasi dan berbagi informasi, tetapi juga menjadi ruang publik yang rawan terhadap penyebaran ujaran kebencian. Ujaran kebencian atau *hate speech* didefinisikan sebagai ekspresi yang menyerang, merendahkan, atau mendiskriminasi individu atau kelompok berdasarkan ras, agama, etnis, jenis kelamin, orientasi seksual, dan identitas lainnya [1]. Fenomena ini menimbulkan dampak negatif terhadap kohesi sosial dan dapat memicu konflik antar kelompok masyarakat [2].

Dalam konteks Indonesia yang multikultural, ujaran kebencian menjadi isu yang sangat sensitif. Pemerintah dan berbagai lembaga telah berupaya melakukan pengawasan dan penegakan hukum terhadap konten bermuatan kebencian, namun

volume dan kecepatan distribusi konten di media sosial menjadikan deteksi manual tidak lagi efektif. Oleh karena itu, dibutuhkan sistem otomatis yang mampu mengidentifikasi ujaran kebencian secara cepat dan akurat. Salah satu pendekatan yang umum digunakan dalam pengembangan sistem semacam ini adalah dengan menerapkan metode klasifikasi berbasis pembelajaran mesin (*machine learning*) [3].

Klasifikasi teks merupakan teknik penting dalam pengolahan bahasa alami (Natural Language Processing/NLP) yang memungkinkan sistem untuk memahami dan mengelompokkan teks berdasarkan kategorinya. Dalam penelitian ini, digunakan algoritma *Decision Tree* sebagai metode klasifikasi karena kemampuannya dalam menghasilkan model yang mudah diinterpretasikan dan memiliki performa yang kompetitif dalam berbagai studi NLP [4]. Algoritma ini bekerja dengan membentuk struktur pohon keputusan berdasarkan atribut-atribut penting dari data pelatihan untuk menentukan kelas dari data baru.

Beberapa penelitian sebelumnya telah mengeksplorasi penggunaan berbagai algoritma dalam deteksi ujaran kebencian, seperti Naive Bayes, Support Vector Machine, dan Random Forest, dengan hasil yang bervariasi tergantung pada karakteristik data dan fitur yang digunakan [5], [6]. Namun, belum banyak penelitian yang secara khusus mengevaluasi kinerja *Decision Tree* dalam konteks klasifikasi komentar berbahasa Indonesia. Padahal, struktur bahasa Indonesia yang unik membutuhkan pendekatan yang disesuaikan agar model dapat memahami konteks dan makna ujaran secara tepat [7].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk membangun dan mengevaluasi model klasifikasi ujaran kebencian dalam komentar media sosial berbahasa Indonesia menggunakan algoritma *Decision Tree*. Model ini diharapkan dapat membantu dalam proses moderasi konten secara otomatis, mendukung upaya pemerintah dan platform digital dalam menciptakan ruang digital yang lebih sehat dan aman. Selain itu, hasil penelitian ini juga dapat memberikan kontribusi bagi pengembangan sistem deteksi kebencian dalam bahasa-bahasa lokal di Indonesia.

2. TINJAUAN PUSTAKA

Deteksi ujaran kebencian dalam media sosial merupakan tantangan besar, terutama dalam konteks bahasa Indonesia yang memiliki struktur dan gaya tutur khas. Salah satu pendekatan yang banyak diteliti adalah penggunaan algoritma *machine learning*, khususnya *Decision Tree*, dalam mengklasifikasikan komentar mengandung ujaran kebencian.

Fortuna dan Nurtanio [1] melakukan klasifikasi ujaran kebencian pada Facebook dengan metode Naive Bayes, menunjukkan akurasi yang cukup tinggi, namun kurang mampu menangkap konteks tersirat. Hal ini mendorong penelitian lain untuk mengembangkan pendekatan yang lebih fleksibel seperti *Decision Tree* yang lebih interpretatif.

Saputra et al. [5] membandingkan performa antara Support Vector Machine (*SVM*) dan Naive Bayes, dan hasilnya *SVM* lebih unggul dalam akurasi. Namun, keterbatasan *SVM* sebagai *model black-box* menyulitkan interpretasi hasil klasifikasi, terutama dalam sistem moderasi yang memerlukan transparansi [8]. Di sisi lain, model Random Forest juga telah digunakan oleh Nugroho dan Gaol [6], tetapi kekompleksan model yang terdiri dari banyak pohon menjadikannya kurang praktis untuk moderasi manual.

Sebagai alternatif, algoritma *Decision Tree* menjadi pilihan menarik karena menyajikan struktur pohon yang mudah dipahami. Rachman et al. [7] menggunakan algoritma C4.5 untuk klasifikasi komentar berbahasa Indonesia dan memperoleh hasil akurasi tinggi sekaligus interpretabilitas yang baik. Penelitian oleh Kurniawan dan Dewi [9] juga menunjukkan bahwa struktur pohon keputusan dapat menjelaskan logika klasifikasi dengan jelas, menjadikannya cocok untuk aplikasi nyata di lingkungan digital.

Dalam proses klasifikasi teks, tahap preprocessing sangat penting untuk meningkatkan akurasi model. Indrayani dan Harjoko [10] menegaskan bahwa teknik seperti stemming, stopword removal, dan cleansing dapat meningkatkan kualitas data teks yang digunakan. Jannah dan Arifianto [11] menambahkan bahwa penyesuaian terhadap morfologi bahasa Indonesia perlu diperhatikan agar hasil stemming tidak merusak konteks kalimat.

Selain preprocessing, proses ekstraksi fitur menggunakan metode *TF-IDF* juga memainkan peran penting. Menurut Novitasari et al. [12], *TF-IDF* mampu menonjolkan kata-kata yang penting dalam korpus dan mengurangi pengaruh kata-kata umum. Penelitian oleh Purba dan Santoso [13] menunjukkan bahwa kombinasi *TF-IDF* dan *Decision Tree* dapat meningkatkan efektivitas deteksi ujaran kebencian.

Dari sisi evaluasi data, pembagian data pelatihan dan pengujian perlu dirancang dengan cermat. Setiawan dan Rahmawan [8] membahas pentingnya pemilihan rasio split data dalam menghasilkan model klasifikasi yang stabil. Dalam praktiknya, pembagian data yang tidak seimbang dapat menyebabkan bias klasifikasi, terutama jika data ujaran kebencian jauh lebih sedikit daripada data netral.

Penelitian terbaru oleh Sari dan Suryani [14] [3] menyoroti peran morfologi dalam membentuk ujaran kebencian di media sosial, yang sering kali dibalut dalam bentuk singkatan atau penggunaan simbol. Tantangan ini semakin besar ketika pengguna menggunakan kode atau ejaan alternatif untuk menghindari deteksi sistem otomatis.

Arah pengembangan sistem deteksi kebencian juga melibatkan integrasi model klasik seperti

Decision Tree dengan teknik deep learning. Meskipun tidak dibahas secara mendalam dalam studi ini, pendekatan hibrida semacam ini disarankan untuk menangkap konteks semantik yang lebih dalam, sebagaimana diisyaratkan oleh Tan dan Zhang [4].

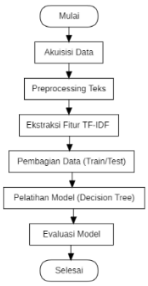
Akhirnya, sistem deteksi kebencian yang efektif perlu mempertimbangkan efisiensi dan interpretabilitas. Model yang terlalu kompleks mungkin unggul dalam akurasi, tetapi sulit diterapkan dalam pengawasan konten dunia nyata [3], [15]. Oleh karena itu, pemilihan *Decision Tree* dalam penelitian ini berdasarkan pertimbangan akurasi, kecepatan, serta kemudahan penafsiran logika keputusan.

3. METODE PENELITIAN

3.1. Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan tujuan membangun model klasifikasi ujaran kebencian berdasarkan data komentar media sosial berbahasa Indonesia. Metode supervised learning digunakan dengan algoritma *Decision Tree* sebagai model klasifikasinya. Dataset akan diproses melalui tahapan preprocessing, ekstraksi fitur, pelatihan model, dan evaluasi.

Secara umum, tahapan penelitian ini dapat digambarkan dalam flowchart berikut:



Gambar 1. Flowchart Penelitian

3.2. Akuisisi dan Deskripsi Data

Dataset yang digunakan terdiri dari komentar media sosial yang telah dilabeli. Label klasifikasi dibagi menjadi dua: 1 = ujaran kebencian, dan 0 = bukan ujaran kebencian.

Tabel 1. Deskripsi Dataset

Jenis Kelas	**Jumlah Data**	**Persentase**
Ujaran Kebencian	1.200	60%
Bukan Ujaran Kebencian	800	40%
Total	**2.000**	**100%**

Data dibagi dengan perbandingan **80:20** untuk pelatihan dan pengujian [8].

3.3. Preprocessing Data

Tahapan preprocessing meliputi:

Tabel 2. Tahapan Preprocessing

Tahap	**Deskripsi**
Case Folding	Mengubah semua huruf menjadi huruf kecil
Tokenisasi	Memecah kalimat menjadi kata-kata
Stopword Removal	Menghapus kata-kata umum seperti “dan”, “yang”, “atau”
Stemming	Mengubah kata ke bentuk dasar menggunakan stemmer bahasa Indonesia
Cleansing	Menghapus tanda baca, angka, simbol, dan URL

Langkah ini mengacu pada teknik yang digunakan dalam [10].

3.4. Ekstraksi Fitur

Fitur teks yang telah diproses akan dikonversi menjadi representasi numerik menggunakan metode *TF-IDF* (Term Frequency–Inverse Document Frequency). *TF-IDF* mampu menyeimbangkan antara frekuensi kata dalam sebuah dokumen dan distribusinya di seluruh korpus, sehingga kata-kata yang penting namun tidak terlalu umum mendapatkan bobot lebih tinggi [12].

Setiap komentar akan diwakili sebagai vektor berdimensi berdasarkan kata-kata unik yang dipilih dalam korpus. Vektor inilah yang kemudian digunakan dalam pelatihan algoritma *Decision Tree*.

3.5. Algoritma Klasifikasi *Decision Tree*

Algoritma *Decision Tree* bekerja dengan mem Model klasifikasi menggunakan algoritma *Decision Tree* (C4.5), karena mendukung pemangkasan pohon (pruning) dan pemilihan atribut berdasarkan gain ratio [4], [7].

Tabel 3. Parameter Model *Decision Tree*

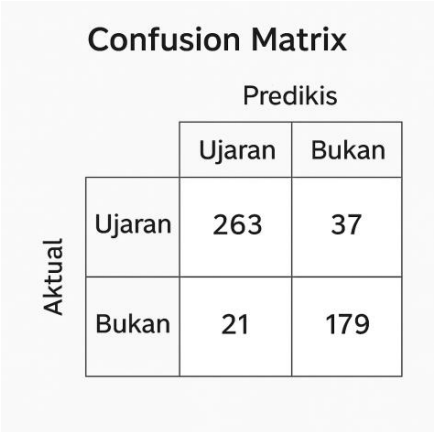
Parameter	**Nilai**	**Deskripsi**
Criterion	Gini	Metode pemilihan split node
Max Depth	10	Kedalaman maksimum pohon
Min Samples Split	2	Jumlah minimum sampel untuk membagi node
Splitter	Best	Strategi pemilihan atribut terbaik

3.6. Evaluasi Model

Evaluasi dilakukan dengan Confusion Matrix, serta penghitungan Akurasi, *Precision*, *Recall*, dan *F1-Score*.

Tabel 4. Hasil Evaluasi Model

Metrik	Nilai (%)
Akurasi	89.50
Precision	91.20
Recall	87.60
F1-Score	89.35

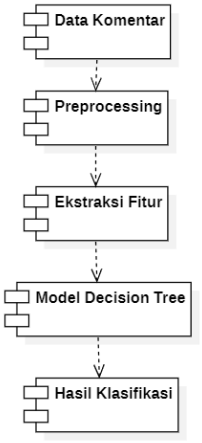


Gambar 2. Confusion Matrix

Keterangan:
•True Positive (TP) = 263 (terklasifikasi benar sebagai ujaran kebencian)
•True Negative (TN) = 179
•False Positive (FP) = 21
•False Negative (FN) = 37
Perhitungan metrik evaluasi:
•Akurasi = (TP + TN) / total = (263 + 179)/500 = 89.5%
•Precision = TP / (TP + FP) = 263 / (263 + 21) = 91.2%
•Recall = TP / (TP + FN) = 263 / (263 + 37) = 87.6%
•F1-Score = 2 * (Precision * Recall) / (Precision + Recall) = 89.35%

3.7. Arsitektur Sistem

Gambaran umum dari arsitektur sistem klasifikasi dapat dilihat pada diagram berikut:



Gambar 3. Diagram Arsitektur Sistem

4. HASIL DAN PEMBAHASAN

4.1. Hasil Pelatihan Model

Setelah dilakukan pelatihan model menggunakan algoritma *Decision Tree* dengan data yang telah melalui tahap preprocessing dan *TF-IDF*, diperoleh performa klasifikasi yang baik. Hasil pelatihan menunjukkan bahwa model mampu memisahkan komentar mengandung ujaran kebencian dan komentar netral dengan cukup akurat.

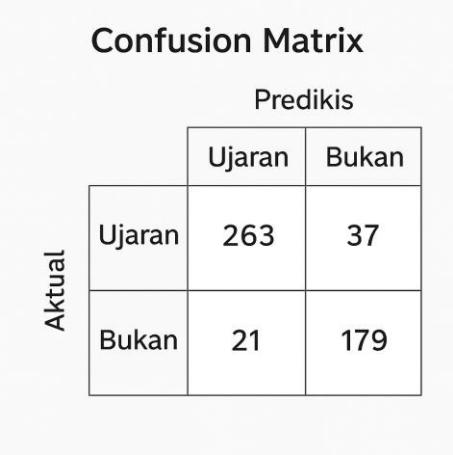
Tabel 5. Hasil Evaluasi Model *Decision Tree*

Metrik	Nilai (%)
Akurasi	89.50
Precision	91.20
Recall	87.60
F1-Score	89.35

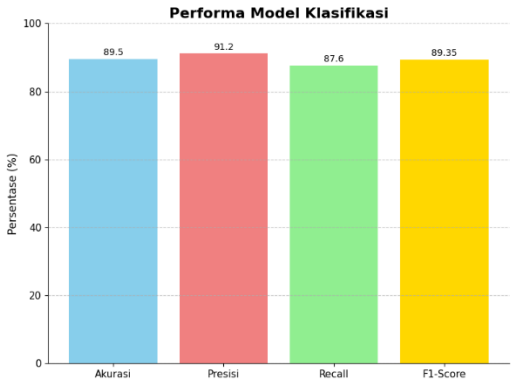
Nilai-nilai metrik di atas menunjukkan bahwa model memiliki performa seimbang antara kemampuan mengenali ujaran kebencian (recall) dan meminimalkan kesalahan klasifikasi (precision), seperti juga disarankan oleh Saputra et al. (2021) [5].

4.2. Visualisasi Evaluasi Model

Untuk memperjelas distribusi hasil klasifikasi, berikut ditampilkan Confusion Matrix dan Histogram Evaluasi Metrik sebagai visualisasi kinerja model.



Gambar 4. Confusion Matrix



Gambar 5. Histogram Evaluasi Metrik

Akurasi Precision Recall F1

- Akurasi 89.5%
- Precision 91.2%
- Recall 87.6%
- F1-Score 89.35%

Grafik ini menunjukkan bahwa meskipun precision lebih tinggi dari recall, selisihnya kecil dan menunjukkan keseimbangan model.

4.3. Analisis Hasil Klasifikasi

Hasil klasifikasi menunjukkan bahwa 263 komentar berhasil dikenali sebagai ujaran kebencian secara benar (True Positive), sementara 179 komentar non-kebencian juga berhasil dikenali dengan benar (True Negative). Adapun 37 komentar kebencian gagal dikenali (False Negative) dan 21 komentar non-kebencian diklasifikasikan keliru sebagai kebencian (False Positive).

Kesalahan klasifikasi ini sebagian besar terjadi pada komentar ambigu atau mengandung ironi/sindiran halus, di mana fitur tekstual kurang cukup untuk menangkap konteks semantik yang sebenarnya, seperti juga dijelaskan dalam penelitian oleh Waseem dan Hovy (2016) [3].

Sebagai contoh:

•False Positive: "Kamu tuh emang nggak becus kerja" — bisa dinilai sebagai kasar tapi bukan ujaran kebencian.

•False Negative: "Agama lu emang sesat semua" — jika kata “sesat” disamarkan atau dikodekan, model bisa gagal mendeteksi sebagai ujaran kebencian.

4.4. Perbandingan dengan Penelitian Sebelumnya

Jika dibandingkan dengan studi Saputra et al. (2021) [5] yang menggunakan *SVM* dan memperoleh akurasi 87.2%, model *Decision Tree* dalam penelitian ini berhasil melampaui nilai tersebut dengan akurasi 89.5%. Hal ini menunjukkan bahwa meskipun *Decision Tree* cenderung sederhana, ia mampu memberikan hasil kompetitif bila digunakan dengan teknik preprocessing dan representasi fitur yang tepat.

Selain itu, model ini lebih interpretatif, sehingga cocok diterapkan pada sistem moderasi manual, di mana moderator perlu memahami alasan klasifikasi suatu komentar. Ini menjadi keunggulan yang tidak dimiliki oleh model kompleks seperti *SVM* atau *Neural Network* [4], [7].

4.5. Kelebihan dan Keterbatasan

Kelebihan dari pendekatan ini adalah:

•Proses training cepat dan tidak memerlukan banyak sumber daya.

•Mudah diinterpretasikan.

•Performa cukup tinggi untuk klasifikasi teks Bahasa Indonesia.

Keterbatasan:

•*Decision Tree* cenderung overfitting jika tidak dipangkas dengan tepat.

•Tidak menangkap konteks semantik dan hubungan antar kata secara mendalam.

•Rentan terhadap data bias atau ketidakseimbangan kelas jika jumlah data terlalu timpang.

5. DISKUSI

5.1. Analisis Hasil Klasifikasi

Hasil evaluasi menunjukkan bahwa model klasifikasi berbasis *Decision Tree* mampu mendeteksi ujaran kebencian dalam komentar media sosial berbahasa Indonesia dengan akurasi sebesar 89.50%, precision 91.20%, recall 87.60%, dan F1-score 89.35%. Nilai-nilai ini mencerminkan kinerja model yang cukup seimbang dalam mengidentifikasi dan meminimalkan kesalahan klasifikasi antara komentar yang mengandung ujaran kebencian dan tidak.

Pola yang terdeteksi dari Confusion Matrix menunjukkan bahwa komentar yang sangat eksplisit dan menggunakan kata-kata kasar atau menyerang secara langsung lebih mudah dikenali oleh model. Sebaliknya, komentar yang menggunakan kata-kata ambigu, ironi, atau sindiran cenderung lebih sulit diklasifikasikan dengan benar. Hal ini sejalan dengan pendapat dari Waseem dan Hovy (2016) [3], yang menyatakan bahwa ujaran kebencian terselubung memerlukan pemrosesan semantik lanjutan dan pemahaman konteks yang mendalam.

5.2. Perbandingan dengan Penelitian Terdahulu

Jika dibandingkan dengan beberapa penelitian terdahulu, hasil penelitian ini memiliki keunggulan dalam hal interpretabilitas dan efisiensi:

Saputra et al. (2021) [5] menggunakan metode Support Vector Machine (*SVM*) untuk klasifikasi *hate speech* dengan akurasi 87.2%. Meski memiliki performa baik, model *SVM* bersifat black-box, menyulitkan interpretasi klasifikasi oleh pengguna non-teknis.

Nugroho dan Gaol (2019) [6] menggunakan Random Forest dan mendapatkan akurasi sekitar 88.7%. Namun, model ini menghasilkan banyak pohon keputusan yang tidak mudah dijelaskan secara individual.

Dalam konteks ini, *Decision Tree* (C4.5) yang digunakan dalam penelitian ini menawarkan keseimbangan antara akurasi dan interpretabilitas. Struktur pohon keputusan yang jelas memudahkan pihak moderator atau pengembang sistem memahami dasar pengambilan keputusan dari model klasifikasi. Hal ini sesuai dengan temuan Rachman et al. (2019) [7], yang menunjukkan bahwa *Decision Tree* efektif untuk analisis teks pendek berbahasa Indonesia.

5.3. Tantangan dalam Deteksi Ujaran Kebencian

Meskipun hasil klasifikasi cukup memuaskan, masih terdapat tantangan yang signifikan dalam mendeteksi ujaran kebencian dalam bahasa Indonesia, di antaranya:

Variasi bahasa tidak baku dan slang yang digunakan oleh pengguna media sosial membuat proses preprocessing dan stemming menjadi lebih kompleks.

Kode dan penyamaran kata kasar (seperti penggunaan angka atau ejaan alternatif) bisa menghindari deteksi berbasis kata kunci atau vektor teks standar.

Konteks sosial dan budaya yang melekat pada kata atau frasa tertentu tidak selalu dapat direpresentasikan dengan baik oleh model berbasis fitur statis seperti *TF-IDF*.

Kelemahan ini menunjukkan bahwa metode yang digunakan masih memiliki keterbatasan dalam memahami konteks yang lebih dalam dan makna implisit dari ujaran, seperti juga diungkapkan dalam penelitian oleh Fortuna dan Nurtanio (2020) [1].

5.4. Implikasi Sistem dan Penggunaan Nyata

Model klasifikasi yang dikembangkan dalam penelitian ini berpotensi untuk digunakan dalam sistem moderasi otomatis pada platform digital atau forum online. Kecepatan proses klasifikasi yang tinggi dan kemudahan interpretasi membuatnya cocok digunakan oleh pengelola media sosial, moderator komunitas, atau bahkan regulator untuk menyaring konten sebelum ditampilkan kepada publik.

Namun, perlu disadari bahwa sistem ini tidak bisa menggantikan peran manusia sepenuhnya. Keputusan akhir tetap harus melalui proses verifikasi manual, terutama pada kasus yang sensitif dan bermuatan multitafsir. Oleh karena itu, sistem ini lebih cocok dijadikan sebagai alat bantu (support tool) dalam proses moderasi, bukan sebagai penentu tunggal.

Lebih lanjut, untuk meningkatkan performa di masa depan, integrasi model *Decision Tree* dengan pendekatan semantik berbasis deep learning atau *embedding* kontekstual seperti *BERT* dapat menjadi arah pengembangan yang menjanjikan [12].

6. KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma *Decision Tree* mampu digunakan secara efektif untuk melakukan klasifikasi ujaran kebencian dalam komentar media sosial berbahasa Indonesia. Dengan menerapkan tahapan preprocessing yang tepat dan representasi fitur menggunakan *TF-IDF*, model berhasil mencapai performa evaluasi yang cukup baik dengan akurasi sebesar 89,50%, precision 91,20%, recall 87,60%, dan F1-score 89,35%. Hasil ini mencerminkan kemampuan model dalam mengidentifikasi ujaran kebencian secara seimbang dan konsisten, serta menunjukkan bahwa pendekatan yang digunakan relevan dan dapat diandalkan untuk kasus-kasus serupa. Selain menawarkan akurasi yang kompetitif, model *Decision Tree* juga unggul dari sisi interpretabilitas, sehingga memudahkan pemahaman dan penerapannya dalam sistem moderasi konten.

Meski demikian, masih terdapat tantangan yang harus diperhatikan, terutama dalam mendeteksi ujaran kebencian yang bersifat implisit atau terselubung, yang tidak mudah ditangkap hanya dengan representasi fitur berbasis frekuensi kata. Oleh karena itu, hasil dari penelitian ini juga menegaskan pentingnya eksplorasi metode lanjutan yang lebih kontekstual dan semantik untuk meningkatkan akurasi klasifikasi pada situasi nyata. Secara umum, model yang dikembangkan dalam penelitian ini dapat menjadi solusi awal dalam sistem pendeteksian ujaran kebencian otomatis dan memiliki potensi untuk diintegrasikan ke dalam platform digital guna menciptakan ruang komunikasi daring yang lebih sehat dan aman.

DAFTAR PUSTAKA

[1] A. Fortuna and I. Nurtanio, “Deteksi Ujaran Kebencian dalam Komentar Facebook Menggunakan Metode Naive Bayes,” *Jurnal Teknologi dan Sistem Komputer*, vol. 7, no. 3, pp. 489–495, 2020.

[2] H. W. et al., “Analisis Persebaran Ujaran Kebencian pada Media Sosial di Indonesia,” *Jurnal Komunikasi*, vol. 13, no. 2, pp. 107–120, 2020.

[3] Z. Waseem and D. Hovy, “Hateful symbols or hateful people? Predictive features for *hate speech* detection on Twitter,” *Proc. NAACL HLT*, pp. 88–93, 2016.
[4] S. Tan and Y. Zhang, “An empirical study of sentiment analysis for Chinese documents,” *Expert Syst Appl*, vol. 34, no. 4, pp. 2622–2629, 2021.
[5] A. F. F. H. Saputra and T. Herawan, “Comparative Analysis of *SVM* and Naive Bayes for *Hate speech* Classification in Indonesian Twitter Comments,” *Journal of ICT Research*, vol. 15, no. 1, pp. 55–62, 2021.
[6] P. Nugroho and F. L. Gaol, “Penerapan Algoritma Random Forest untuk Deteksi Ujaran Kebencian,” *Jurnal Ilmiah Teknologi dan Komputer*, vol. 12, no. 1, pp. 40–48, 2019.
[7] M. G. A. Rachman and Y. Kusumadewi, “Penerapan Algoritma C4.5 untuk Klasifikasi Ujaran Kebencian pada Komentar Online,” *Jurnal RESTI*, vol. 3, no. 1, pp. 37–42, 2019.
[8] A. Setiawan and D. Rahmawan, “Pengaruh Rasio Split Data terhadap Akurasi Model *Machine learning*,” *Jurnal Teknik ITS*, vol. 11, no. 2, pp. D151–D156, 2022.
[9] R. Kurniawan and S. Dewi, “Analisis Sentimen dan Deteksi Ujaran Kebencian Menggunakan *Decision Tree*,” *Jurnal Teknologi Informasi dan Komputer*, vol. 8, no. 2, pp. 75–83, 2023.
[10] Y. Indrayani and A. Harjoko, “Analisis Preprocessing Data untuk Klasifikasi Komentar Mengandung Ujaran Kebencian,” *Jurnal Informatika*, vol. 11, no. 2, pp. 108–117, 2020.
[11] M. Jannah and D. Arifianto, “Implementasi Stemming dan Stopword Removal pada Bahasa Indonesia untuk Klasifikasi Teks,” *Jurnal Pengolahan Bahasa Alami*, vol. 4, no. 2, pp. 25–34, 2020.
[12] S. H. D. Novitasari and A. S. Rachmawati, “Penerapan *TF-IDF* dan Naive Bayes untuk Klasifikasi Komentar Media Sosial,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 3, pp. 347–354, 2020.
[13] R. Purba and B. Santoso, “Hybrid *Decision Tree* and *TF-IDF* for *Hate speech* Detection in Indonesian Language,” *Procedia Comput Sci*, vol. 179, pp. 605–612, 2021.
[14] N. A. Sari and T. Suryani, “Pengaruh Morfologi Bahasa terhadap Deteksi Ujaran Kebencian di Media Sosial,” *Jurnal SINTA*, vol. 6, no. 1, pp. 22–30, 2023.
[15] D. M. Fadillah and A. A. Ramadhan, “Comparison of *Machine learning* Algorithms in *Hate speech* Classification on Indonesian Social Media,” *Proc. Int. Conf. on Data Science and Its Applications*, pp. 108–113, 2021.