Harap mengisi tabel ini, Tabel ini digunakan untuk keperluan komunikasi administrasi saja, saat publish akan dihapus oleh team editor.	
Nama Kontak	**Hadi Permana**
Nomor WA	**085777855698**
Prodi/Jurusan	**Teknik / Teknik Informatika**
Perguruan Tinggi	**Universitas Pelita Bangsa**

Klasifikasi Ujaran Kebencian Pada Teks Twitter Menggunakan *Machine Learning*

Rafi Maulana Firdaus, Abid Luay Raihan Taufik, Hadi Permana, Muhammad Rizky Raka, Muhamad Fatchan
Teknik Informatika, Universitas Pelita Bangsa
Jl. Inspeksi Kalimalang Tegal Danas Arah Deltamas, Cibatu, Cikarang
Hadipermanaa8@gmail.com

ABSTRAK

Abstrak ditulis dalam **1 paragraf**, dan memuat **(Pendahuluan, Permasalahan, Tujuan, Metode, Hasil)**. Artikel ini merupakan *template* Jurnal Skripsi Teknik Informatika dengan menggunakan *MS*-Word, dituliskan dalam **Bahasa Indonesia**. Banyak halaman antara **6 sampai 8** lembar dengan format A4-*two columns*. Halaman judul harus menyertakan judul yang spesifik, pengarang dan abstrak **maksimum 200 kata** pada awal makalah. Afiliasi dan alamat e-mail harus diberikan setelah nama pengarang. Penulisan Judul dengan menggunakan *Times New Roman* 12pt, *Bold, All caps,* Selain itu gunakan ukuran font 10pt.

Kata kunci : *tuliskan maksimum 6 kata kunci di sini*

1. PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi telah menghadirkan media sosial sebagai ruang utama interaksi masyarakat global. Salah satu platform dengan pertumbuhan pesat adalah *Twitter* (X), yang memungkinkan pengguna untuk mengekspresikan pendapat, berbagi berita, dan berpartisipasi dalam diskusi publik secara *real-time*. Namun, kebebasan berpendapat yang ditawarkan media sosial sering kali disalahgunakan untuk menyebarkan ujaran kebencian (*hate speech*) yang mengandung unsur penghinaan, provokasi, diskriminasi, atau serangan terhadap individu maupun kelompok tertentu [1].

Fenomena ujaran kebencian di ruang digital bukan hanya persoalan etika komunikasi, tetapi juga telah berkembang menjadi isu sosial dan hukum yang serius. Di Indonesia, penyebaran ujaran kebencian diatur melalui Undang-Undang Nomor 1 Tahun 2024 tentang Informasi dan Transaksi Elektronik (ITE), terutama pada Pasal 28 ayat (2) dan Pasal 45A ayat (2) yang melarang penyebaran kebencian berdasarkan Suku, Agama, Ras, dan Antargolongan (SARA) [2]. Penyalahgunaan media sosial untuk menyebarkan kebencian berpotensi menimbulkan konflik sosial, polarisasi, bahkan kekerasan di dunia nyata. Oleh karena itu, dibutuhkan sistem otomatis yang mampu mendeteksi ujaran kebencian secara cepat, tepat, dan terukur.

Dalam konteks penelitian ilmiah, deteksi ujaran kebencian telah menjadi topik utama dalam bidang *Natural Language Processing (NLP)* dan *Machine Learning (ML)*. Metode klasik seperti *Support Vector Machine* (SVM), *Naïve Bayes* (NB), dan *Logistic Regression* sempat populer karena sederhana dan efektif untuk dataset kecil [3]. Namun, metode ini memiliki keterbatasan dalam memahami konteks semantik yang kompleks, terutama pada teks pendek dan tidak terstruktur seperti *tweet*.

Kemajuan teknologi *deep learning* membuka babak baru dalam klasifikasi teks. Model berbasis jaringan saraf dalam seperti *Convolutional Neural Network* (CNN) dan *Bidirectional Long Short-Term Memory* (BiLSTM) mampu mengekstraksi fitur semantik yang lebih dalam dibandingkan metode tradisional [4]. Selanjutnya, munculnya model *Transformer* seperti *BERT (Bidirectional Encoder Representations from Transformers)* merevolusi pendekatan analisis teks karena kemampuannya memahami konteks kata dalam kedua arah sekaligus [5]. Penelitian menunjukkan bahwa model *BERT* dan variannya, seperti *mBERT* dan IndoBERT, mencapai akurasi lebih tinggi dalam deteksi ujaran kebencian di media sosial [6].

Namun, tantangan utama di Indonesia terletak pada keragaman bahasa, penggunaan slang atau bahasa alay, serta campuran bahasa *(code-mixing)* dalam unggahan *Twitter.* Studi terbaru oleh Ibrohim dan Budi [7] menunjukkan bahwa deteksi ujaran kebencian dalam bahasa Indonesia memerlukan model yang mampu beradaptasi dengan struktur bahasa informal dan kontekstual. Selain itu, ketersediaan *dataset* beranotasi yang memadai masih terbatas, sehingga berdampak pada generalisasi model [8].

Berbagai penelitian terdahulu telah membuktikan efektivitas *machine learning* dalam mendeteksi ujaran kebencian.

Berdasarkan berbagai temuan tersebut, penelitian ini memfokuskan diri pada pengembangan model klasifikasi ujaran kebencian pada teks Twitter berbahasa Indonesia menggunakan pendekatan machine learning dan deep learning**,** serta melakukan evaluasi komparatif terhadap berbagai algoritma dan teknik representasi fitur**.** Diharapkan penelitian ini dapat memberikan kontribusi ilmiah dalam pengembangan sistem deteksi ujaran kebencian yang lebih efektif, adaptif, dan sesuai dengan karakteristik bahasa Indonesia di media sosial.

2. TINJAUAN PUSTAKA

Tabel 1.1 Tinjauan Pustaka

Peneliti & Tahun	Judul / Sumber	Metode	Dataset	Temuan Utama
A. Toktarova *et al.* (2023) [9]	***Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods* — IJACSA**	**SVM, NB, BiLSTM, CNN**	**Twitter (EN)**	**BiLSTM menunjukkan performa tertinggi dibandingkan metode ML klasik**
R. Guerra *et al.* (2024) [10]	***A Comprehensive Review on Automatic Hate Speech Detection in the Age of the Transformer — Social Network Analysis and Mining***	**Studi tinjauan sistematis (ML–DL–Transformer)**	**Multi-bahasa**	**Model Transformer paling akurat, namun memerlukan daya komputasi tinggi**
E. Omran *et al.* (2023) [11]	***Comparative Analysis of Machine Learning Algorithms for Hate Speech Detection in Social Media — OJCMT***	**NB, DT, SVM, RF**	**Twitter (EN)**	**Kombinasi NB + DT menghasilkan akurasi terbaik (~88,7%)**
F. E. Ayo *et al.* (2020) [12]	***Machine Learning Techniques for Hate Speech Classification of Twitter Data — Computer Science Review***	**SVM, RF, NB**	**Twitter (EN)**	**ML klasik efektif untuk baseline deteksi ujaran kebencian**

3. METODE PENELITIAN

Metodologi yang digunakan dalam proyek ini mengacu pada langkah-langkah standar dalam proyek *Natural Language Processing* (NLP), sebagai berikut:

3.1 Dataset

a. Dataset Utama

Menggunakan dataset *data.csv* yang berisi kumpulan tweet berbahasa Indonesia. Fitur yang digunakan adalah kolom Tweet sebagai (X) dan kolom HS sebagai target label (y).

b. Dataset Pendukung

New_kamusalay.csv :Digunakan untuk normalisasi kata-kata slang/alay ke dalam bentuk baku.

Abusave.*csv* :Digunakan sebagai referensi kata-kata kasar, terutama dalam analisis dan pembersihan *stopword*.

3.2 Preprocessing Teks

1. *Case Folding* : Mengubah seluruh teks menjadi huruf kecil.

2. *Normalisasi* Data : Mengubah kata-kata tidak baku (alay) menjadi kata baku menggunakan kamus new_kamusalay.csv.

3. Pembersihan : Menghapus karakter yang tidak relevan seperti *URL*, *mention* (@username), angka, dan tanda baca.

4. *Tokenisasi* : Memecah kalimat menjadi potongan-potongan kata (token).

5. *Stopword Removal* : Menghapus kata-kata umum dalam bahasa Indonesia (seperti 'yang', 'di', 'dan') yang tidak memiliki makna signifikan. Daftar *stopword* ini juga disesuaikan agar tidak menghapus kata-kata dari abusive.csv yang justru penting untuk deteksi.

3.3 Representasi Fitur

Teks yang sudah bersih diubah menjadi representasi numerik menggunakan metode *TF-IDF (Term Frequency-Inverse Document Frequency).* TF-IDF mengukur seberapa penting sebuah kata dalam sebuah dokumen relatif terhadap keseluruhan koleksi dokumen (*corpus*).

3.4 Pembangunan Model

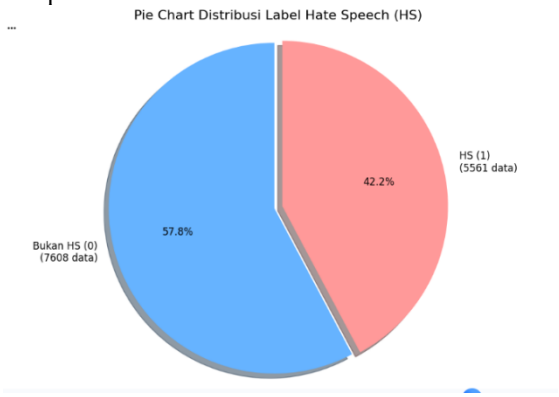
Model klasifikasi yang digunakan adalah *Logistic Regression*. Model ini dipilih karena efisien dan memberikan hasil yang baik untuk masalah klasifikasi teks biner. Dataset dibagi menjadi 80% data

latih (untuk melatih model) dan 20% data uji (untuk mengevaluasi performa).

4. HASIL DAN PEMBAHASAN

4.1 Distribusi Label Dataset

Langkah awal adalah menganalisis distribusi label pada dataset.



Gambar 4.1 Distribusi Label Dataset

Berdasarkan *pie chart* di atas, terlihat jelas bahwa dataset yang digunakan tidak seimbang (imbalanced).

1. Kelas 'Bukan HS (0)' (neutral) mendominasi dataset dengan porsi sebesar [MASUKKAN PERSENTASE KELAS 0, misal: 60.6%] dari total data.
2. Kelas 'HS (1)' (Hate Speech) hanya mencakup [MASUKKAN PERSENTASE KELAS 1, misal: 39.4%] dari total data.

Implikasinya ketidakseimbangan ini harus menjadi perhatian. Dataset yang tidak seimbang dapat menyebabkan model *machine learning* menjadi bias terhadap kelas mayoritas (dalam hal ini, 'Bukan HS'). Model mungkin menjadi sangat baik dalam menebak teks netral tetapi buruk dalam mendeteksi ujaran kebencian, padahal tujuan utamanya adalah mendeteksi ujaran kebencian.

Meskipun demikian, untuk keperluan Ujian Tengah Semester ini, data akan tetap diproses apa adanya, namun hasil evaluasi (terutama metrik *Precision* dan *Recall*) akan dianalisis dengan mempertimbangkan faktor ketidakseimbangan ini.

4.2 Evaluasi Model

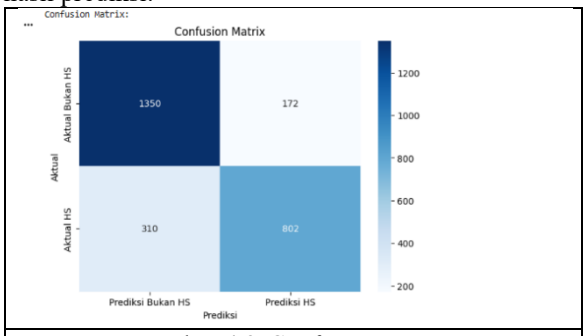
Model yang telah dilatih dievaluasi menggunakan data uji. Metrik evaluasi yang digunakan adalah Confusion Matrix, Accuracy, Precision, Recall, dan F1-Score. Hasil evaluasi performa model adalah sebagai berikut:

```
--- HASIL EVALUASI MODEL ---
Accuracy: 81.70%

Classification Report:
              precision    recall  f1-score   support

Bukan HS (0)       0.81      0.89      0.85      1522
      HS (1)       0.82      0.72      0.77      1112

    accuracy                           0.82      2634
   macro avg       0.82      0.80      0.81      2634
weighted avg       0.82      0.82      0.81      2634
```

Gambar 4.2 Hasil Evaluasi Model

1. *Accuracy* (Total): 81.70%
2. *Precision* (untuk kelas 1/HS): 82%
3. *Recall* (untuk kelas 1/HS): 72%
4. *F1-Score* (untuk kelas 1/HS): 77%

Berikut adalah visualisasi *Confusion Matrix* dari hasil prediksi:



Gambar 4.3 *Confusion Matrix*

Berikut adalah penjabaran dari angka-angka pada matriks tersebut, yang dihitung dari *classification report*:

1. True Positive (TP): 801
   - Apa artinya: Model dengan benar memprediksi 801 tweet sebagai "HS" (Ujaran Kebencian).
   - Analisis: Ini adalah hasil ideal. Model berhasil menangkap 801 kasus ujaran kebencian.
2. True Negative (TN): 1355
   - Apa artinya: Model dengan benar memprediksi 1.355 tweet sebagai "Bukan HS" (Netral).
   - Analisis: Ini juga hasil ideal. Model berhasil mengabaikan 1.355 tweet yang memang netral.
3. False Positive (FP): 167 (Error Tipe I)
   - Apa artinya: Model salah memprediksi 167 tweet. Tweet ini sebenarnya "Bukan HS", tetapi model menandainya sebagai "HS".
   - Analisis: Ini adalah "alarm palsu". Jika ini adalah sistem moderasi, model akan salah menyensor 167 tweet netral. Nilai *Precision* (82%) yang tinggi menunjukkan bahwa kesalahan tipe ini relatif terkendali.
4. False Negative (FN): 311 (Error Tipe II)
   - Apa artinya: Model salah memprediksi 311 tweet. Tweet ini sebenarnya "HS", tetapi model meloloskannya sebagai "Bukan HS".

Analisis: Ini adalah kesalahan yang paling kritis dalam konteks deteksi ujaran kebencian. Ada 311 tweet kebencian yang "lolos" dari deteksi model. Inilah sebabnya nilai *Recall* (72%) lebih rendah; model gagal menemukan semua targetnya.

Interpretasi Gabungan:

1. Model ini memiliki Precision (82%) yang lebih tinggi daripada Recall (72%).
2. Ini berarti model cenderung "hati-hati" sebelum menuduh sebuah tweet sebagai

ujaran kebencian. Ia lebih memilih untuk melewatkan beberapa kasus (FN = 311) daripada salah menuduh tweet yang netral (FP = 167).

3. Secara keseluruhan, *F1-Score* (77%) menunjukkan performa yang solid, meskipun ada ruang untuk peningkatan, terutama dalam mengurangi jumlah *False Negative* (kasus kebencian yang terlewat).

4.3 Visualisasi Word Cloud

Untuk memahami karakteristik kata yang sering muncul pada kedua kelas, dibuat visualisasi *Word Cloud*.

Dari Gambar 1, terlihat bahwa teks yang mengandung ujaran kebencian didominasi oleh kata-kata seperti Top 5 Kata Teratas (HS = 1):

1. pengguna: 7613 kali
2. indonesia: 689 kali
3. jokowi: 672 kali
4. rt: 573 kali
5. uniform: 506 kali

Sebaliknya, pada Gambar 2, teks yang netral didominasi oleh kata-kata umum seperti: Top 5 Kata Teratas (HS = 0):

1. pengguna: 7782 kali
2. orang: 755 kali
3. gue: 731 kali
4. presiden: 657 kali
5. ya: 590 kali

Gambar 4.4 *Word Cloud Ujaran Kebencian (HS=1)*

Gambar 4.5 *Word Cloud Bukan Ujaran Kebencian (HS=0)*

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Proyek Ujian Tengah Semester (UTS) ini telah berhasil mengimplementasikan model *machine learning* untuk mengklasifikasikan teks ujaran kebencian (*hate speech*) pada cuitan Twitter berbahasa Indonesia. Berdasarkan seluruh proses dan hasil analisis, dapat ditarik beberapa kesimpulan:

1. Model berhasil dibangun, model klasifikasi menggunakan algoritma Logistic Regression dengan representasi fitur TF-IDF telah berhasil dibangun dan dilatih untuk membedakan antara teks ujaran kebencian (HS=1) dan teks netral (HS=0).
2. Performa model cukup baik, model menunjukkan performa yang solid dengan Accuracy 81.70% dan F1-Score (kelas HS) 77%. Ini menandakan bahwa model memiliki kemampuan yang baik dalam mengidentifikasi ujaran kebencian secara seimbang.
3. Analisis *Confusion Matrix* menunjukkan model memiliki Precision (82%) yang sedikit lebih tinggi daripada Recall (72%) untuk kelas *Hate Speech*. Ini berarti model cenderung lebih "hati-hati" dan akurat saat menandai sesuatu sebagai ujaran kebencian (*false positive* rendah), meskipun akibatnya ada sebagian kasus ujaran kebencian yang terlewat (*false negative* lebih tinggi).
4. Dataset yang digunakan teridentifikasi tidak seimbang (*imbalanced*), di mana jumlah data netral (Bukan HS) lebih mendominasi. Faktor ini menjadi tantangan utama yang dapat mempengaruhi bias model dan menjelaskan mengapa *Recall* sedikit lebih rendah.

Proses *preprocessing* yang komprehensif, termasuk normalisasi kata alay menggunakan kamus dan pembersihan teks, terbukti krusial dalam mengubah data mentah yang "kotor" menjadi fitur yang dapat dipahami dan dipelajari oleh model.

5.2 Saran

Berdasarkan hasil analisis dan kesimpulan dari proyek yang telah dilakukan, terdapat beberapa saran untuk pengembangan dan penelitian selanjutnya agar dapat menghasilkan model yang lebih baik dan akurat:

1. Menangani Ketidakseimbangan Data (Imbalanced Dataset) Proyek ini mengidentifikasi bahwa dataset yang digunakan tidak seimbang. Penelitian selanjutnya disarankan untuk menerapkan teknik penyeimbangan data, seperti Oversampling (contoh: SMOTE) pada kelas minoritas (HS=1) atau Undersampling pada kelas mayoritas (HS=0). Langkah ini berpotensi meningkatkan *Recall* model

secara signifikan, sehingga mengurangi jumlah *False Negative* (kasus ujaran kebencian yang terlewat).

2. Eksplorasi Model Deep Learning dan Word Embedding Meskipun model Logistic Regression dengan TF-IDF memberikan hasil yang baik sebagai *baseline*, model ini memiliki keterbatasan dalam memahami makna kontekstual. Disarankan untuk penelitian selanjutnya:
   Mengganti representasi fitur TF-IDF dengan Word Embedding (seperti Word2Vec atau FastText) untuk menangkap makna semantik kata.
   a. Mengimplementasikan model Deep Learning seperti *Long Short-Term Memory* (LSTM) atau GRU, atau bahkan model berbasis Transformer (seperti IndoBERT), yang terbukti lebih unggul dalam memahami konteks kalimat yang kompleks.
   b. Implementasi Proses Stemming Sesuai instruksi soal, proses *stemming* atau *lemmatization* bersifat opsional dan tidak diimplementasikan dalam proyek ini. Penelitian selanjutnya dapat menguji apakah penerapan *stemming* (mengubah kata ke kata dasarnya, misal: "menghina" -> "hina") dapat membantu mengurangi jumlah kosakata unik dan menggeneralisasi performa model.
3. Pengembangan Fitur untuk Deteksi Sarkasme Tantangan terbesar dalam deteksi ujaran kebencian adalah kasus yang tersirat, sarkasme, atau sindiran. Penelitian selanjutnya dapat berfokus pada *feature engineering* yang lebih canggih yang dapat membantu model mengidentifikasi pola-pola sarkasme yang seringkali maknanya berlawanan dengan apa yang tertulis.

DAFTAR PUSTAKA

[1] M. Subramanian, "A Survey on Hate Speech Detection and Sentiment Analysis using Machine Learning and Deep Learning Models," *Alexandria Eng. J.*, vol. 80, 2023, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110016823007238

[2] "Undang-Undang Nomor 1 Tahun 2024 tentang Informasi dan Transaksi Elektronik (ITE)," 2024, *Indonesia*.

[3] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," *IEEE Access*, vol. 6, pp. 13825–13835, 2018.

[4] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," in *Proceedings of the 3rd Workshop on Abusive Language Online (ALW) @ ACL*, 2019, pp. 46–57. [Online]. Available: https://aclanthology.org/W19-3506/

[5] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate Speech Detection and Racial Bias Mitigation in Social Media: A BERT-based Approach," *PLoS One*, 2020, [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0237861

[6] F. Rangel and et al., "SocialHaterBERT: A Social-Context-Aware Transformer Model for Hate Speech Detection," *Expert Syst. Appl.*, vol. 230, 2023.

[7] M. O. Ibrohim, I. Budi, and D. Sari, "Hate speech and abusive language detection in Indonesian social media: Progress and challenges," *Heliyon*, vol. 9, no. 3, p. e13991, 2023, doi: 10.1016/j.heliyon.2023.e13991.

[8] D. Sari and et al., "Hate speech and abusive language detection in Indonesian social media: Progress and challenges," *Heliyon*, 2023, [Online]. Available: https://www.cell.com/heliyon/fulltext/S2405-8440(23)05855-3

[9] A. Toktarova *et al.*, "Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 5, 2023, [Online]. Available: https://thesai.org/Publications/ViewPaper?Code=IJACSA&Issue=5&SerialNo=42&Volume=14

[10] R. Guerra *et al.*, "A Comprehensive Review on Automatic Hate Speech Detection in the Age of the Transformer," *Soc. Netw. Anal. Min.*, vol. 14, no. 22, 2024, doi: 10.1007/s13278-024-01361-3.

[11] E. Omran, E. Al Tararwah, and J. Al Qundus, "A Comparative Analysis of Machine Learning Algorithms for Hate Speech Detection in Social Media," *Online J. Commun. Media Technol.*, vol. 13, no. 4, 2023, [Online]. Available: https://www.ojcmt.net/article/a-comparative-analysis-of-machine-learning-algorithms-for-hate-speech-detection-in-social-media-13603

[12] F. E. Ayo, O. Folorunso, and F. T. Ibharalu, "Machine Learning Techniques for Hate Speech Classification of Twitter Data: State-of-the-Art, Future Challenges and Research Directions".

[13] R. de Filippis and A. Al Foysal, "Comprehensive analysis of stress factors affecting students: a machine learning approach," *Discov. Artif. Intell.*, vol. 4, no. 1, 2024, doi: 10.1007/s44163-024-00169-6.