Available *online* at: https://bit.ly/SISFOTEKJAR

Jurnal Sistem Informasi dan Teknologi Jaringan

(SISFOTEKJAR)

ISSN (Print) 2808-7917 | ISSN (Online) 2807-9256

Penggunaan Deep Learning untuk Mengklasifikasi Hate speech dan Good Speech Terhadap Pertamina di Platform Twitter dengan Metode Convolutional Neural Network (CNN)

Hasan A. Situmorang[1], Martiano[2]

[1,2] Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia
[1]situmoranghasan@gmail.com, [2]martiano@umsu.ac.id

ARTICLE INFORMATION

Received: June 27, 2025
Revised: Agust 22, 2025
Available online: September 08, 2025

KEYWORDS

Deep Learning, CNN, Hate Speech, Good Speech, Twitter

CORRESPONDENCE

Phone: +62 851-7110-9412

E-mail: Situmoranghasan@gmail.com

ABSTRACT

Advances in digital technology have changed the way people express their opinions, particularly through social media platforms like Twitter. In social and corporate contexts, Pertamina, a state-owned energy company in Indonesia, is frequently the subject of public discourse, both in the form of positive (good speech) and negative (hate speech) expressions. To manage this information, a system capable of accurately and automatically classifying tweets is crucial. This study aims to develop a Deep Learning-based text classification model, specifically using the Convolutional Neural Network (CNN) method, to identify tweets containing hate speech and good speech related to Pertamina. Data was collected from Twitter using relevant keywords, followed by manual preprocessing and labeling. The cleaned dataset was then divided into training and testing data for processing using a CNN architecture. The results showed that the CNN model performed very well in the classification task, achieving a validation accuracy of 98.08% and a testing accuracy of 97.66%. Evaluation using a confusion matrix also showed high precision and recall values, with an f1 score of 0.98. These findings demonstrate that the CNN method is effective in accurately identifying hate speech and positive speech in Indonesian text data, particularly regarding issues related to Pertamina. This research is expected to contribute to the development of automated social media monitoring systems and public opinion management tools.

1. PENDAHULUAN

Media sosial kini menjadi ruang interaksi publik yang bebas dan terbuka, namun juga memiliki potensi besar dalam menyebarkan ujaran kebencian (*hate speech*) [1]. Fenomena *hate speech* dapat memberikan dampak negatif, baik secara sosial maupun ekonomi, seperti menurunkan citra perusahaan dan memicu keresahan di masyarakat [2]. Pertamina sebagai salah satu Badan Usaha Milik Negara (BUMN) strategis di bidang energi sering kali menjadi objek percakapan publik di *platform Twitter*, baik dalam bentuk kritik maupun apresiasi. Oleh karena itu, diperlukan suatu pendekatan cerdas untuk mengidentifikasi dan menganalisis sentimen publik terhadap Pertamina secara otomatis dan akurat.

Berbagai penelitian sebelumnya telah menunjukkan bahwa *Convolutional Neural Network* (CNN) mampu memberikan akurasi tinggi dalam tugas-tugas deteksi ujaran kebencian dan analisis sentimen [3]. CNN terbukti lebih unggul dibandingkan metode tradisional seperti *Naïve Bayes* dan *Support Vector Machine* (SVM) yang umumnya hanya mengandalkan analisis berbasis kata tanpa memperhatikan konteks semantik. Keunggulan CNN terletak pada kemampuannya mengenali pola linguistik yang kompleks serta efisiensinya dalam memproses data dalam jumlah besar [4]. Selain itu, CNN memiliki kecepatan pelatihan yang lebih tinggi dibandingkan metode berbasis urutan

seperti *Long Short-Term Memory* (LSTM), sehingga sangat sesuai diterapkan pada dataset tweet yang besar dan dinamis [5].

Dalam penelitian ini, pendekatan CNN dimanfaatkan untuk mengklasifikasikan tweet tentang Pertamina ke dalam dua kategori, yaitu *hate speech* dan *good speech*. Dengan penerapan model CNN, penelitian ini diharapkan dapat membantu dalam mengidentifikasi opini publik secara otomatis berdasarkan konten yang diunggah di media sosial. Pendekatan ini tidak hanya memberikan kontribusi terhadap pengembangan ilmu pengetahuan di bidang kecerdasan buatan (*artificial intelligence*) dan pemrosesan bahasa alami (*natural language processing*), tetapi juga memiliki nilai praktis yang signifikan bagi institusi seperti Pertamina dalam memantau persepsi publik secara real-time.

Tujuan dari penelitian ini adalah: (1) membangun model klasifikasi tweet menjadi hate speech dan good speech menggunakan algoritma CNN; (2) mengintegrasikan teknik CNN untuk menangkap pola-pola semantik dalam teks guna meningkatkan kinerja klasifikasi; dan (3) mengukur tingkat akurasi metode CNN dalam mengklasifikasikan hate speech dan good speech pada data tweet mengenai Pertamina. Dengan pencapaian tujuan tersebut, diharapkan hasil penelitian ini dapat menjadi dasar bagi pengembangan sistem analisis sentimen otomatis yang lebih efektif dalam mendukung pengambilan keputusan strategis berbasis opini publik.

2. TINJAUAN PUSTAKA

1. Media Sosial Twitter

Twitter adalah layanan jejaring sosial berbasis teks singkat (tweet) yang dapat digunakan untuk berbagi opini, informasi, bahkan sebagai alat kampanye dan kritik sosial secara terbuka dan cepat [6]. Twitter merupakan salah satu platform media sosial yang berfokus pada penyebaran pesan singkat atau tweet, yang awalnya dibatasi hanya sampai 140 karakter dan kini telah berkembang hingga 280 karakter [7]. Platform ini memungkinkan penggunanya untuk membagikan pemikiran, informasi, opini, maupun respons terhadap isu-isu tertentu secara cepat dan terbuka. Dengan sifatnya yang real-time dan publik, Twitter menjadi salah satu media sosial yang paling banyak digunakan dalam diskusi sosial, politik, dan ekonomi, termasuk dalam menanggapi kebijakan pemerintah atau aktivitas perusahaan negara seperti Pertamina.

2. Hate Speech Dan Good Speech

Ujaran kebencian (*Hate Speeech*) adalah bentuk ekspresi yang menyampaikan penghinaan terhadap kelompok tertentu yang dianggap inferior, dan sering kali digunakan sebagai alat dominasi oleh kelompok mayoritas [8]. Sedangkan Good Speech adalah ujaran yang mengandung nilai positif seperti pujian, motivasi, apresiasi, atau dukungan terhadap individu maupun kelompok tertentu [9].

Klasifikasi antara *hate speech* dan *good speech* menjadi penting untuk memahami persepsi masyarakat secara utuh terhadap Pertamina di media sosial, khususnya Twitter. Melalui pendekatan berbasis teknologi, seperti deep learning menggunakan metode *Convolutional Neural Network* (CNN), proses deteksi dan klasifikasi ujaran ini dapat dilakukan secara otomatis dan efisien. Analisis semacam ini tidak hanya membantu dalam pemetaan opini publik, tetapi juga menjadi dasar bagi perusahaan untuk mengambil langkah komunikasi yang lebih strategis dan responsif.

3. Convolutional Neural Network

Convolutional Neural Network (CNN) merupakan salah satu arsitektur *deep learning* yang awalnya dikembangkan untuk pengolahan citra (*image processing*), namun dalam perkembangannya telah terbukti efektif juga dalam menangani data teks, termasuk dalam tugas klasifikasi seperti deteksi hate speech dan good speech [10]. CNN bekerja dengan cara mengekstraksi fitur dari data masukan menggunakan lapisan konvolusi (*convolutional layer*), yang kemudian dilanjutkan dengan proses pooling dan klasifikasi.

Dalam konteks pengolahan teks, CNN mampu mendeteksi pola-pola linguistik, frasa, atau kombinasi kata tertentu yang menjadi indikator dari suatu kelas tertentu, misalnya ujaran kebencian atau ujaran positif. Keunggulan utama CNN adalah kemampuannya dalam melakukan ekstraksi fitur secara otomatis dari teks mentah tanpa memerlukan pemrograman manual untuk menentukan fitur atau kata kunci tertentu. Hal ini sangat bermanfaat dalam analisis data Twitter yang memiliki beragam variasi bahasa, gaya penulisan, serta penggunaan bahasa informal atau tidak baku.

3. METODE PENELITIAN

Metode penelitian ini menggunakan pendekatan kuantitatif eksperimental yang memanfaatkan deep learning untuk melakukan klasifikasi teks, dengan fokus pada deteksi otomatis ujaran kebencian (*hate speech*) dan ujaran positif (*good speech*) terhadap Pertamina di platform media sosial Twitter. Pendekatan kuantitatif dipilih karena data yang dianalisis berupa data numerik dan teks yang telah diproses secara digital, sehingga hasilnya dapat diukur menggunakan metrik evaluasi seperti akurasi, presisi, dan recall. Penelitian bersifat eksperimental karena melibatkan proses perancangan, pelatihan, dan pengujian model kecerdasan buatan melalui serangkaian eksperimen terhadap data yang telah dikumpulkan sebelumnya.

Selain itu, penelitian ini menggunakan pendekatan komputasional karena seluruh proses klasifikasi dilakukan dengan bantuan algoritma *Convolutional Neural Network* (CNN), yang dikenal efektif dalam memproses data teks dan mengenali pola bahasa yang kompleks, termasuk dalam konteks informal seperti tweet di media sosial. Berdasarkan metodologi yang telah dirancang, penelitian dilaksanakan melalui beberapa tahapan yang terstruktur dan saling berkesinambungan, mulai dari penentuan fokus penelitian, pengumpulan data, pengolahan data, hingga penarikan kesimpulan, sehingga diperoleh hasil yang akurat, relevan, dan dapat dipertanggungjawabkan secara ilmiah. Adapun langkah-langkah penelitian yang dilakukan adalah sebagai berikut:

1. Dataset

Pada tahap ini, data dikumpulkan dari platform media sosial Twitter dengan fokus pada percakapan publik yang menyebutkan Pertamina dan BBM (Bahan Bakar Minyak), yang relevan dengan topik penelitian. Pengumpulan data bertujuan untuk mendapatkan tweet yang mengandung opini publik, baik dalam bentuk hate speech (ujaran kebencian) maupun good speech (ujaran positif atau dukungan).

Dataset terdiri dari 5.980 tweet yang dikumpulkan dengan kata kunci terkait Pertamina. Dari total data, 80% digunakan sebagai data latih (training) dan 20% sebagai data validasi (testing). Distribusi dataset sangat tidak seimbang, dengan jumlah good speech lebih banyak dibanding hate speech.

2. *Processing Data*

Setelah data terkumpul, dilakukan tahap pra-pemrosesan untuk membersihkan data teks dari elemen-elemen yang tidak relevan. Teks pada tweet dikonversi ke huruf kecil, kemudian dilakukan penghapusan karakter khusus, tautan (URL), simbol "@" pada mention, tagar (hashtag), angka, serta tanda baca yang tidak diperlukan. Proses ini juga mencakup penghapusan stopword, yakni kata-kata umum yang tidak memiliki kontribusi besar terhadap makna teks, serta tokenisasi, yaitu memecah kalimat menjadi kata-kata terpisah untuk mempermudah pemetaan selanjutnya dalam bentuk angka.

3. *Labeling*

Langkah berikutnya adalah pelabelan data. Karena tidak tersedia label langsung dalam dataset mentah, peneliti menyusun daftar kata kasar, sindiran, dan sarkasme yang relevan dengan konteks Pertamina dan isu sosial di sekitarnya. Dengan bantuan daftar tersebut, tweet dikategorikan ke dalam dua kelas, yaitu hate speech dan good speech. Proses ini dilakukan secara semi-otomatis. Label ditentukan dengan mendeteksi keberadaan kata kasar/ujaran kebencian. Tweet dengan kata kasar diberi label 1 (hate speech), sedangkan yang lainnya diberi label 0 (good speech).

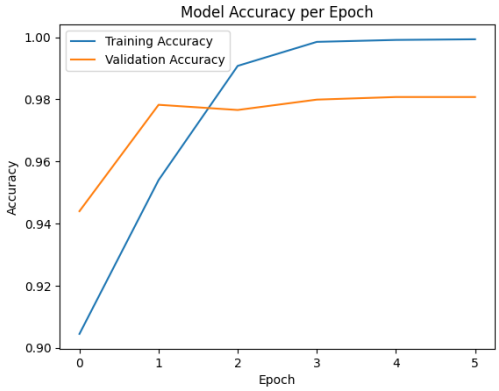
4. *Model Cnn*

Pada tahap berikutnya, dibangun model CNN untuk klasifikasi teks dengan arsitektur yang dirancang secara sistematis. Model ini diawali dengan embedding layer yang berfungsi mengubah token atau kata menjadi vektor representasi numerik. Selanjutnya, lapisan konvolusi (Conv1D) digunakan untuk mendeteksi pola-pola lokal atau n-gram yang bermakna dalam teks. Hasil dari lapisan konvolusi kemudian diproses oleh GlobalMaxPooling1D guna mengekstraksi fitur-fitur terpenting dan mereduksi dimensi output. Fitur-fitur yang telah diperoleh diteruskan ke dense layer dengan 64 unit dan fungsi aktivasi ReLU untuk memproses hasil ekstraksi, sementara dropout layer dengan tingkat dropout 0.5 disisipkan di antara lapisan fully connected untuk mencegah overfitting. Terakhir, output layer dengan fungsi aktivasi sigmoid digunakan untuk menghasilkan prediksi biner, yaitu 0 untuk good speech dan 1 untuk hate speech. Model ini dilatih menggunakan 80% data latih dari total dataset, sedangkan 20% sisanya digunakan untuk pengujian guna mengevaluasi performa klasifikasi.

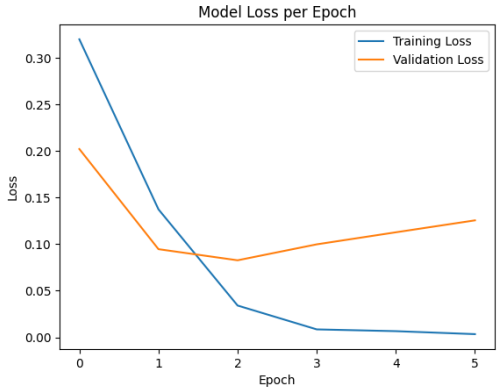
4. HASIL DAN PEMBAHASAN

1. *Hasil Training Dan Validasi Model*

Akurasi data pelatihan pada grafik akurasi model per epoch meningkat secara signifikan dari 90,4% pada epoch ke-0 menjadi 99,9% pada epoch ke-5. Akurasi validasi juga meningkat cukup konsisten, dari 94,5% pada epoch awal hingga stabil di kisaran 98,0% pada epoch ke-3 hingga ke-5. Kecenderungan stabilitas akurasi validasi menunjukkan bahwa model tidak mengalami overfitting yang signifikan, karena tidak terjadi penurunan akurasi pada data validasi walaupun akurasi pelatihan meningkat drastis. Stabilitas grafik validasi menunjukkan bahwa model memiliki kemampuan generalisasi yang baik terhadap data yang belum pernah dilihat sebelumnya (data validasi).



Gambar 1. Grafik Akurasi CNN



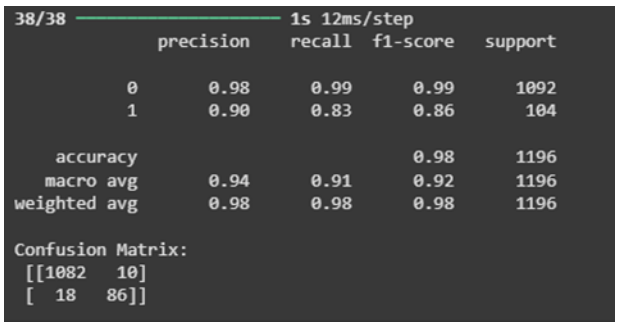
Gambar 2. Grafik Loss Model CNN

Loss data pelatihan menurun tajam dari sekitar 0.32 pada epoch ke-0 menjadi mendekati 0.00 pada epoch ke-5, yang menandakan model sangat cocok dengan data latih. Loss validasi juga mengalami penurunan hingga mencapai nilai minimum pada epoch ke-2 (~0.08), tetapi kemudian sedikit meningkat kembali hingga ~0.13 pada epoch ke-5.Kenaikan kecil pada validation loss setelah epoch ke-2 dapat mengindikasikan tanda awal overfitting, meskipun masih tergolong ringan dan tidak menyebabkan degradasi signifikan terhadap akurasi

2. *Hasil Evalusi Model*

Hasil evaluasi menunjukkan Akurasi model sebesar 98%, artinya 98% tweet berhasil diklasifikasikan dengan benar. Precision Good Speech (label 0) sangat tinggi (0.98), menandakan hampir semua prediksi good speech benar. Precision Hate Speech (label 1) sebesar 0.90, artinya 90% tweet yang diprediksi sebagai hate speech memang benar. Recall untuk Hate Speech hanya 0.83, artinya masih terdapat 17% tweet hate speech yang tidak berhasil dikenali (false negative).

F1-Score untuk Hate Speech (0.86) cukup baik, tetapi menunjukkan bahwa model sedikit lebih kesulitan mengenali ujaran kebencian dibanding good speech. Macro average (rata-rata sederhana dari kedua kelas) F1-Score adalah 0.92, dan weighted average juga 0.98, memperlihatkan bahwa model tetap sangat andal secara keseluruhan meskipun distribusi data tidak seimbang. Dari 1.092 tweet good speech, sebanyak 1.082 berhasil diklasifikasikan dengan benar, dan hanya 10 salah diklasifikasikan sebagai hate speech.Dari 104 tweet hate speech, 86 diklasifikasikan dengan benar, sementara 18 salah diklasifikasikan sebagai good speech. Dengan jumlah data hate speech yang jauh lebih sedikit, model tetap mampu mendeteksi sebagian besar kasus ujaran kebencian dengan akurat.



```
38/38 ━━━━━━━━━━━━━━━━━━━━ 1s 12ms/step
              precision    recall  f1-score   support

           0       0.98      0.99      0.99      1092
           1       0.90      0.83      0.86       104

    accuracy                           0.98      1196
   macro avg       0.94      0.91      0.92      1196
weighted avg       0.98      0.98      0.98      1196

Confusion Matrix:
 [[1082   10]
 [  18   86]]
```

Gambar 3. Hasil Evaluasi Model

3. Hasil Klasifikasi Data

Hasil klasifikasi menunjukkan bahwa model berhasil mengidentifikasi sebanyak 1.092 tweet sebagai good speech dan 104 tweet sebagai hate speech pada data validasi, yang menandakan bahwa mayoritas percakapan terkait Pertamina di Twitter cenderung bersifat positif.

No	Text	Label_Asli	Prediksi
281	kontribusi pertamina bikin bangga bangsa indonesia	0	1
49	apresiasi buat pertamina yang jaga pasokan energi nasional	0	0
8	pertamina terus dukung umkm berkembang lewat berbagai program	0	1
101	pertamina jelas gagal jaga kestabilan harga	1	1
192	janji palsu pertamina soal subsidi rakyat kecil	1	1
55	rakyat makin susah gara-gara ulah pertamina	1	1
187	pelayanan spbu makin buruk, gak ada perubahan	1	1
123	bangga punya pertamina sebagai perusahaan energi nasional	0	0
251	pertamina emang gak pernah transparan soal kenaikan harga bbm	1	0
136	harga bbm terus naik, rakyat kecil makin tertekan	1	1
95	spbu bikin masalah tiap libur panjang	1	0
76	pertamina emang gak pernah transparan soal kenaikan harga bbm	1	0
134	apresiasi buat pertamina yang jaga pasokan energi nasional	0	0
109	masyarakat dirugikan, bbm mahal tapi kualitas rendah	1	0
16	pertamina terus dukung umkm berkembang lewat berbagai program	0	0
294	janji palsu pertamina soal subsidi rakyat kecil	1	1
287	keren, distribusi bbm tetap lancar berkat pertamina	0	1
244	salut, pertamina dukung energi ramah lingkungan	0	1
199	spbu sering kosong, pelayanan bikin kecewa	1	0

Gambar 4 Hasi Klasifikasi Data

5. KESIMPULAN

Model CNN yang dikembangkan menunjukkan kinerja klasifikasi yang sangat baik dengan akurasi sebesar 98%, nilai precision sebesar 0.98, recall sebesar 0.98, dan f1-score sebesar 0.98 secara keseluruhan (weighted average). Berdasarkan hasil classification report, model mampu mendeteksi tweet kategori good speech dengan precision sebesar 0.98 dan recall sebesar 0.99, serta kategori hate speech dengan precision sebesar 0.90 dan recall sebesar 0.83. Hal ini menunjukkan bahwa model cukup seimbang dalam mendeteksi kedua jenis ujaran, meskipun performa untuk hate speech sedikit lebih rendah karena kompleksitas dan variasi ekspresinya. Hasil confusion matrix juga memperlihatkan bahwa dari total 1.196 tweet, model berhasil mengklasifikasikan 1.082 tweet good speech secara benar dari 1.092 tweet dan 86 tweet hate speech secara benar dari 104 tweet. Terdapat 10 false positive (good speech diklasifikasikan sebagai hate speech) dan 18 false negative (hate speech diklasifikasikan sebagai good speech).

Meskipun penelitian ini hanya berfokus pada opini publik terhadap Pertamina, peneliti selanjutnya disarankan untuk menerapkan metode serupa pada objek lain seperti tokoh publik, instansi pemerintah, atau isu yang sedang viral agar dapat diterapkan secara lebih luas dalam analisis sentimen media sosial. Selain itu, karena klasifikasi saat ini masih terbatas pada dua kelas, yaitu hate speech dan good speech, disarankan agar penelitian berikutnya memperluas klasifikasi menjadi multi-kelas seperti kritik membangun, provokatif, atau sarkastik guna menghasilkan analisis opini publik yang lebih mendalam dan komprehensif.

DAFTAR PUSTAKA

[1] A. Rizkiyah, M. Mayasari, and T. W. Budhiharti, "Ujaran Kebencian (Hate Speech) Pada Public Figure Dalam Kolom Komentar Di Media Sosial," *J. Ilm. Wahana Pendidik.*, vol. 10, no. 20, pp. 87–95, 2024.

[2] S. Supiyandi, S. Khodijah, N. S. Sitha, Y. Sembiring, and N. R. Fauzan, "Tinjauan Dampak Negatif Fenomena Kebencian di Media Sosial di Indonesia," *Senashtek 2024*, vol. 2, no. 1, pp. 77–80, 2024.

[3] I. Pangestuti, "KLASIFIKASI KOMENTAR ABUSIVE TEKS TWITTER MENGGUNAKAN METODE CONVOLUTIONAL NEURAL NETWORK," in *Seminar Nasional Teknoka 7*, Jurnal Uhamka, 2022.

[4] Y. Helsa, *Artificial Intelligence untuk Pendidikan Strategi Pembelajaran, Efisiensi Guru, dan Implementasi Mengajar AI untuk Siswa di Setiap Level*. Deepublish, 2025.

[5] M. W. S. Utomo, H. W. Murti, A. W. I. Sujatmoko, and A. P. Sari, "DETEKSI SPAM EMAIL MENGGUNAKAN METODE LSTM (LONG SHORT TERM MEMORY)," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 6, pp. 11406–11411, 2024.

[6] Z. P. Ajie and A. Wardhana, "Analisis Sentimen Negatif Publik pada Tagar# KaburAjaDulu di Media Sosial 'X,'" *J. Audiens*, vol. 6, no. 3, pp. 430–442, 2025.

[7] M. A. Muqsith, *Pesan politik di media sosial Twitter*. Jakad Media Publishing, 2022.

[8] A. Rahmadhany, A. A. Safitri, and I. Irwansyah, "Fenomena penyebaran hoax dan hate speech pada media sosial," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 3, no. 1, pp. 30–43, 2021.

[9] R. L. Hasanah and E. H. Hermaliani, "Perbandingan Tradisional dan Ensemble Machine Learning dalam Melakukan Klasifikasi Kalimat Ujaran Kebencian," *Insect (Informatics Secur. J. Tek. Inform.*, vol. 8, no. 2, pp. 121–131, 2023.

[10] S. Hidayat, Y. V. Via, and E. P. Mandyartha, "Penerapan Model Hybrid Convolutional Neural Network dan Long Short-Term Memory untuk Pengenalan Real-Time Sistem Isyarat Bahasa Indonesia (SIBI)," *J. MEDIA Inform. BUDIDARMA*, vol. 8, no. 3, pp. 1586–1596, 2024.