

Klasifikasi Sentimen Komentar Forum Reddit menggunakan Fine-Tuned IndoBERT (Studi Kasus: Program Makan Bergizi Gratis)

Oleh:

Bram Aji Saka Putra,
Dr. Suprianto S.Si., M.Si

Progam Studi

Universitas Muhammadiyah Sidoarjo

Februari, 2026



Pendahuluan

Perkembangan internet telah mengubah cara masyarakat berkomunikasi dan memperoleh informasi melalui media sosial. Reddit dipilih sebagai sumber data karena memungkinkan teks panjang, struktur subreddit terorganisir, anonimitas tinggi, dan tersedianya API untuk ekstraksi data. Program Makan Bergizi Gratis diangkat sebagai fokus penelitian karena merupakan kebijakan strategis nasional dengan implementasi berkelanjutan. Model IndoBERT dipilih karena merupakan model bahasa yang telah dioptimalkan khusus untuk bahasa Indonesia dengan kemampuan embeddings kontekstual yang mampu menangani variasi gaya tulisan, tingkat formalitas, penggunaan bahasa sehari-hari, serta kompleksitas linguistik dan nuansa sentimen yang kaya karakteristik pada data media sosial Reddit.

Pertanyaan Penelitian (Rumusan Masalah)

Penelitian ini mencoba menjawab sejauh mana akurasi model IndoBERT dalam mengklasifikasikan sentimen, serta bagaimana distribusi sentimen publik terhadap Program Makan Bergizi Gratis, serta apa saja karakteristik linguistik yang membedakan masing-masing kategori sentimen.

Metode

Penelitian menggunakan pendekatan yang terdiri dari enam fase berurutan. Tahapan dimulai dari pengumpulan data menggunakan PRAW API, dilanjutkan dengan preprocessing teks meliputi pembersihan dan normalisasi, persiapan data melalui pelabelan manual dengan pembagian data latih dan uji, proses fine-tuning model IndoBERT dengan optimasi hyperparameter, evaluasi menggunakan metrik standar, serta prediksi dan analisis pada keseluruhan dataset.

Hasil

Dari proses pengumpulan data, diperoleh 6.295 komentar final setelah penyaringan. Model yang dikembangkan mencapai akurasi 99% dengan F1-score, precision, dan recall yang sama tingginya. Kecepatan pemrosesan model mencapai 137,93 sampel per detik. Hasil prediksi menunjukkan distribusi sentimen netral 59,44% persen, positif 34,11%, dan negatif 6.45 %. Rata-rata confidence score yang diperoleh adalah 0,8502.

Pembahasan

Model menunjukkan konvergensi yang cepat dan stabil tanpa indikasi overfitting dengan hanya 6 kesalahan klasifikasi dari 600 data uji. Dominasi sentimen netral mengindikasikan karakteristik diskusi di Reddit yang bersifat informatif dan analitis daripada ekspresi opini eksplisit. Confidence score tertinggi diperoleh pada kategori netral, sedangkan sentimen negatif menunjukkan score terendah karena tantangan dalam mendeteksi sarkasme dan kritik implisit.

Temuan Penting Penelitian

Penelitian berhasil membuktikan keunggulan fine-tuning IndoBERT dengan konfigurasi optimal yang menghasilkan akurasi tinggi. Platform Reddit terbukti efektif sebagai sumber data karena memfasilitasi diskusi mendalam dengan karakteristik linguistik yang kaya. Masing-masing kategori sentimen menunjukkan pola komunikasi yang berbeda, di mana sentimen positif cenderung elaboratif, negatif bersifat ringkas dan emosional, serta netral menggunakan terminologi teknis. Keterbatasan utama teridentifikasi pada kemampuan model dalam mendeteksi sentimen negatif yang tersirat.

Manfaat Penelitian

Secara teoritis, penelitian memperkaya literatur pengolahan bahasa alami untuk bahasa Indonesia dan memberikan landasan untuk penelitian longitudinal serupa. Secara praktis, hasil penelitian dapat dimanfaatkan pemerintah untuk memantau respons publik secara real-time, peneliti untuk mengembangkan metodologi serupa pada domain lain, serta praktisi komunikasi dalam merumuskan strategi berbasis insight sentimen masyarakat.

Referensi

- B. Auxier and M. Anderson, "Social Media Use in 2021," Pew Research Center, 2021. [Online]. Available: www.pewresearch.org.
- K. Pham, K. C. Rao Kathala, and S. Palakurthi, "Reddit Sentiment Analysis on the Impact of AI Using VADER, TextBlob, and BERT," *Procedia Computer Science*, vol. 258, pp. 85-94, 2025.
- A. Kiftiyah et al., "Program Makan Bergizi Gratis (MBG) dalam Perspektif Keadilan Sosial dan Dinamika Sosial-Politik," *Pancasila: Jurnal Keindonesiaan*, vol. 5, no. 1, pp. 45-62, 2025.
- R. A. Munir, "Analisis Sentimen Cuitan di Media Sosial X tentang Program Makan Bergizi Gratis dengan Metode NLP," *Jurnal Informatika Dan Teknik Elektro Terapan*, vol. 13, no. 1, pp. 123-135, 2025.
- V. Agustina, A. Herliana, and E. P. Korespondensi, "Analisis Sentimen Publik atas Kebijakan Efisiensi Anggaran 2025 dengan Text Mining dan Natural Language Processing," *Jurnal Sistem Informasi dan Teknologi*, vol. 7, no. 2, pp. 78-89, 2025.
- B. Boe, "PRAW: The Python Reddit API Wrapper," GitHub Documentation, 2023. [Online]. Available: <https://praw.readthedocs.io/en/stable/index.html>

Referensi

- M. Rayhan Nur, Y. Wibisono, and R. Megasari, "Analisis Sentimen dan Pemodelan Topik pada Post tentang Merek Teknologi di X Menggunakan Fine-tuning IndoBERT dan BERTopic," Jurnal Komputer Teknologi Informasi Sistem Informasi (JUKTISI), vol. 4, no. 2, pp. 45-58, 2025.
- Babanejad, N., Agrawal, A., An, A., & Papagelis, M. (2020). A Comprehensive Analysis of Preprocessing for Word Representation Learning in Affective Tasks. <https://github.com/NastaranBa/>
- M. Faza and M. Taufik, "Penerapan Model IndoBERT Untuk Deteksi Potensi Sumber Stres Dalam Teks Media," Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 12, no. 3, pp. 567-578, 2025.
- A. Kunaefi, Z. Abidin, and R. Kusumawati, "Klasifikasi Berita Hoaks Bahasa Indonesia Menggunakan IndoBERT Fine-Tuning Dengan Pendekatan Focal Loss pada Data Tidak Seimbang," JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika), vol. 10, no. 2, pp. 89-102, 2025.
- D. Prasetia et al., "Analisis Sentimen Pengguna Aplikasi MyBluebird Dengan Algoritma Naïve Bayes Di Playstore," Jurnal Informatika Dan Teknik Elektro Terapan, vol. 13, no. 2, pp. 145-156, 2025.
- N. Proferes et al., "Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics," Social Media and Society, vol. 7, no. 2, pp. 1-14, 2021.

Referensi

- W. Wang et al., "IndoBERT: Pre-trained Model for Bahasa Indonesia," IndoNLP Research Group, 2020. [Online]. Available: <https://huggingface.co/indobenchmark>
- Z. Liu et al., "Improving Sentiment Analysis Accuracy with Emoji Embedding," Journal of Safety Science and Resilience, vol. 2, no. 4, pp. 289-298, 2021.
- M. Arif Afandi and I. Suri, "Effectiveness of Government Communication in Delivering Public Policy on Social Media," Jurnal Komunikasi Pemerintah, vol. 9, no. 1, pp. 23-38, 2025.
- N. Ratnaswari, N. C. Wibowo, and D. S. Y. Kartika, "Analisis Sentimen Menggunakan Metode Lexicon-Based Dan Support Vector Machine pada Presiden dan Wakil Presiden Indonesia Periode 2024–2029," Jurnal Informatika Dan Teknik Elektro Terapan, vol. 13, no. 1, pp. 67-78, 2025.
- S. S. Berutu et al., "Data preprocessing approach for machine learning-based sentiment classification," JURNAL INFOTEL, vol. 15, no. 4, pp. 317-325, 2023.
- H. T. Duong and T. A. Nguyen-Thi, "Preprocessing techniques and data augmentation for sentiment analysis," Computational Social Networks, vol. 8, no. 1, pp. 1-15, 2021.

Referensi

- R. Srinivasan and C. N. Subalalitha, "Sentimental analysis from imbalanced code-mixed data using machine learning approaches," Distributed and Parallel Databases, vol. 41, no. 1-2, pp. 37-52, 2023.
- A. Ari Putra Wibowo and Nurul Hidayat, "Eksplorasi Linguistik Komputasional dalam Analisis Bahasa Alami untuk Mengungkap Evolusi Dialek Digital di Era Media Sosial Global," Journal of New Trends in Sciences, vol. 1, no. 3, pp. 45-52, 2023.
- Nurjoko and A. Rahardi, "Model Indo-BERT untuk Identifikasi Sentimen Kekerasan Verbal di Twitter," IJCCS (Indonesian Journal of Computer and Cybernetics), vol. 18, no. 2, pp. 156-168, 2024.
- F. A. Wagay and Jahiruddin, "Classification of Mental Illnesses from Reddit Posts Using Sentence-BERT Embeddings and Neural Networks," Procedia Computer Science, vol. 258, pp. 234-243, 2025.

