e-ISSN: 3165-1453 https://doi.org/10.36782/gitcoin.v1i1.692

Journal of Digital Transformation, Computing, and Intelligence (GITCOIN)
Vol. 1, No. 1, Maret 2026

# Analisis Sentimen Publik terhadap Program Makan Bergizi Gratis pada TikTok Menggunakan XGBoost dan IndoBERT

**Muhammad Rafi Ibnusalis Nurzaman[1)*], Faliza Dineta[1)], Dimas Aryo Anggoro[1)]**

[1)] Department of Informatics, Faculty of Enginering and Computer Science, Universitas Bakrie, Jakarta, Indonesia
*rafiibnusalisnurzaman@gmail.com, fnetaa066@gmail.com, dimas.anggoro@bakrie.ac.id

**ABSTRAK**

TikTok adalah *platform* media sosial yang paling digunakan secara aktif dari seluruh kalangan umur, menjadi salah satu tempat diskusi publik, salah satunya terkait program Makan Bergizi Gratis (MBG). Opini masyarakat tersebar luas di *platform* media sosial ini dengan bentuk teks dari hasil komentar, yang memiliki bahasa tidak terstruktur dan penuh dengan bahasa gaul. Yang berpengaruh terhadap sulitnya penerimaan masukan secara otomatis, sementara metode survei membutuhkan waktu yang lama. Penelitian ini bertujuan untuk mengukur tingkat penerimaan masyarakat terhadap Makan Bergizi Gratis dengan klasifikasi data teks komentar pada *platform* TikTok ke dalam tiga variabel keputusan sentimen. Dengan total data teks komentar sebanyak 57.362 yang kemudian akan dibersihkan dengan *confidence filtering*, memperoleh data valid sebesar 42.023 komentar. Menggunakan RoBERTa sebagai pelabelan otomatis dan *pre-trained* model IndoBERT untuk mentransformasikan teks bersih menjadi metrik, kemudian menggunakan algoritma *supervised learning* XGBoost untuk klasifikasi akhir. Hasil pengujian *Hybrid* IndoBERT-XGBoost memiliki nilai *Accuracy* 95,21% dengan *F1-Score* sebesar 95,18%. Temuan ini dapat memberikan kontribusi dalam bentuk *prototype dashboard* sebagai produk akhir yang berfungsi untuk memantau akumulasi volume sentimen publik secara *real-time*

***Kata kunci:*** IndoBERT, XGBoost, *Data Mining*, MBG, TikTok

***Riwayat artikel:*** Diterima 1 Juli 2026, Direvisi 21 Juli 2026, Diterima 3 Agustus 2026

***ABSTRACT***

*TikTok is a widely used social media platform across all age groups, serving as a venue for public discourse—including discussions regarding the Free Nutritious Meal (MBG) program. Public opinion is abundant on the platform in the form of comments characterized by unstructured language and heavy use of slang. This poses challenges for automated input processing, whereas traditional survey methods are time-consuming. This study aims to measure public acceptance of the Free Nutritious Meal program by classifying TikTok comment data into three sentiment categories. Out of 57,362 raw comments, 42,023 valid entries were obtained after applying Confidence Filtering. The methodology employed RoBERTa for automated labeling and the pre-trained IndoBERT model to transform cleaned text into metrics, followed by the XGBoost supervised learning algorithm for final classification. Testing of the Hybrid IndoBERT-XGBoost model yielded an accuracy of 95.21% and an F1-score of 95.18%. These findings contribute to the development of a dashboard prototype designed to aggregate public sentiment volume in real-time.*

***Keywords:*** IndoBERT, XGBoost, *Data Mining*, MBG, TikTok

***Article History:*** Accepted 1 July 2026, Revised 21 July 2026, Accepted 3 Agustus 2026

## 1. PENDAHULUAN

Media sosial menjadi salah satu tempat masyarakat untuk menyampaikan pendapat mereka tentang berbagai topik. TikTok adalah *platform* media sosial yang paling sering aktif digunakan sebanyak 2 miliar pengguna dari berbagai kalangan umur, yang salah satunya digunakan sebagai ruang opini publik di seluruh dunia [12]. TikTok dapat digunakan untuk menyampaikan kritik dan dukungan melalui komentar pada suatu konten video, yang dapat menjadi pusat perbincangan serta menghasilkan topik yang viral.[3]

Secara demografis, TikTok didominasi oleh kalangan generasi muda. Sekitar 38,5% dari total pengguna adalah kelompok usia 18 hingga 24 tahun. *Platform* media sosial ini mengungguli *platform* besar lainnya seperti

© 2026 Journal of Digital Transformation, Computing, and Intelligence (GIT-COIN)
This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY-NC 4.0).

Youtube yang memiliki persentase penggunaan 23% dan Instagram yang memiliki persentase penggunaan 22%, dan memiliki tingkat engagement terbesar di mana pengguna TikTok rata-rata menghabiskan 95 menit per harinya.[13]

Salah satu topik yang menjadi tempat berkumpulnya kritik dan dukungan adalah program Makan Bergizi Gratis [19]. Topik ini ramai diperbincangkan masyarakat Indonesia sebagai salah satu janji politik utama pada masa pemilihan umum presiden.[20] Karena beragamnya opini publik yang tersebar di *platform* media sosial, penerapan analisis sentimen menjadi sangat penting.[6] Analisis sentimen bertujuan untuk memahami opini publik terhadap sebuah topik berdasarkan data acak yang telah disiapkan dari internet atau media sosial, yang akan diolah dalam skala luas untuk mengidentifikasi persepsi masyarakat.[4]

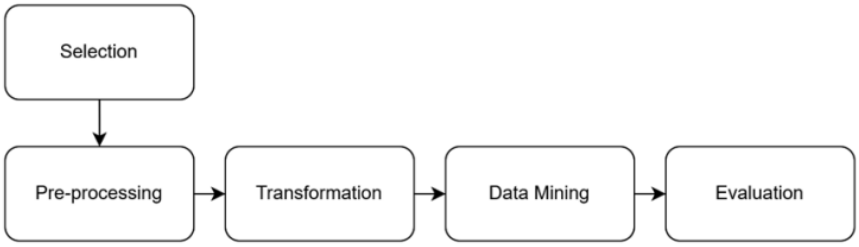
Namun, permasalahannya ada di gaya bahasa masyarakat pada kolom komentar video TikTok yang tidak terstruktur dan berbahasa gaul, yang menjadi masalah untuk mengevaluasi kebijakan secara cepat dan otomatis, dan membutuhkan waktu yang lama jika melakukan dengan cara survei satu per satu.[1] IndoBERT adalah varian *pre-trained* BERT yang dilatih dengan bahasa Indonesia, yang mampu untuk mentranslasikan bahasa tidak baku dan gaya bahasa gaul, menjadi bahasa yang terstruktur dan dapat diolah dalam skala besar.[4] Di sisi lain, algoritma XGBoost digunakan sebagai algoritma berbasis *Decision Trees* yang dapat mengolah data dalam jumlah besar (*Big Data*).[2] Model dan algoritma ini dapat menghasilkan nilai *Accuracy* 95,21% dengan *F1-Score* sebesar 95,18% pada data uji, dari *dataset* awal sebanyak 57.362 komentar, akan dibersihkan dengan *confidence filtering*, memperoleh data valid sebesar 42.023 komentar.[9]

Oleh karena itu, fokus utama penelitian ini adalah untuk menguji efektivitas model IndoBERT dan XGBoost dalam analisis sentimen 3 kelas (positif, negatif, netral) pada topik Makan Bergizi Gratis (MBG).[2] Hubungan penelitian ini dengan yang terdahulu adalah penelitian ini secara langsung menguji model yang jauh lebih modern dengan data yang lebih banyak. Penelitian ini berkontribusi dalam membentuk *prototype dashboard* sebagai produk akhir yang berfungsi untuk memantau akumulasi volume sentimen publik secara *real-time*.[3]

## 2. METODE

### 2.1. Tahapan Penelitian

Penelitian ini menggunakan metode penelitian kuantitatif, yang menerapkan algoritma *Hybrid* IndoBERT dan XGBoost untuk melakukan analisis sentimen opini masyarakat terhadap Makan Bergizi Gratis, dengan menggunakan *Knowledge Discovery in Databases* yang melalui 5 proses utama yaitu:[11]




**Gambar 1.** *Knowledge Discovery in Databases*

Gambar 1 adalah proses *Knowledge Discovery in Databases* yang terdapat *selection*, *pre-processing*, *transformation*, *data mining*, dan *evaluation*. Dimulai dari mengambil *dataset* utama, mengeksplorasi karakteristik data, membersihkan data teks dari *special characters*, *pseudo labeling*, dan membagi *dataset* menjadi *training data* dan *testing data* dengan metode *stratified sampling*, kemudian hasilnya akan dievaluasi untuk menilai akurasi dalam pengujian performa.

### 2.2. Selection

*Dataset* yang digunakan bersumber dari *platform* Kaggle dengan judul "*Dataset* Komentar TikTok MBG (Makan Bergizi Gratis)" https://www.kaggle.com/datasets/sinryurifal/*dataset*-komentar-tiktok-mbg-makan-bergizi-gratis/data. *Dataset* ini terdiri dari 8 *class*, namun hanya *class "comment"* yang dibutuhkan, sehingga menghasilkan 57.362 data komentar.

**Tabel 1.** *Dropped Dataset*

| Sample | Comment |
|---|---|
| 1 | ibuk ngomongnya, TIMES NEW ROMAN, BOLD, CAPLOCK... |
| 2 | ENTER ENTER ENTER Langsung ke INTI nya |
| 3 | SHIFT+ENTER langsung ke lembar berikutnya 🗿 |
| 4 | @Lfajeriah |
| 5 | 🤣🤣🤣😭😭 |

Tabel 1 menampilkan 5 data pertama dari *dataset* yang semua *class*-nya dihapus kecuali *class "comment"* yang menjadi objek yang akan diolah dalam penelitian ini, yang terdiri atas tiga kelas analisis sentimen (positif, negatif, netral). Seluruh sampel akan digunakan untuk menganalisis opini publik terhadap Makan Bergizi Gratis. Tahap ini juga didukung dengan proses analisis eksplorasi data, yang dilakukan untuk memahami distribusi dan deviasi struktur pada *dataset* sebelum data diproses.

Proses analisis eksplorasi data mencakup pengujian terhadap, panjang karakter (*character length)*, jumlah kata *(word count)*, *noise analysis* yang mengidentifikasi data-data kotor pada *dataset*, seperti komentar khusus dengan *emoji*, *mention*, dan tautan. Lalu terakhir adalah dengan mengidentifikasi kata-kata populer untuk 30 kata *most frequent words*, yang akan dibersihkan dari *dataset* untuk menangkap tujuan awal topik (contoh: "di", "yang", "nya", "mbg", "makan", "kebijakan").

### 2.3. Pre-processing

Proses dimulai dengan mengubah seluruh karakter huruf di dalam *dataset* menjadi huruf kecil, dan *text cleaning* untuk menghapus seluruh tautan, *mention*, numerik, *special characters*, dan *emoji*.[13] Dilanjutkan dengan proses *pseudo labeling* yang memberikan pelabelan otomatis terhadap *raw dataset*, dengan memanfaatkan model arsitektur *pre-trained* RoBERTa Bahasa Indonesia. Proses ini mencakup *label encoding* yang menentukan *label encoder* untuk mengubah kategori data menjadi numerik (negatif = 0, netral = 1, positif = 2).[5] Dan *confidence filtering* untuk memvalidasi *pseudo label,* dengan menerapkan batas skor keyakinan (*confidence threshold*) sebesar $\geq$ 0,80. Untuk data yang memiliki skor di bawah 0,80 akan dihapus dari *dataset*. Proses penyaringan ini akan memberikan data bersih yang akan digunakan untuk tahap pemodelan. [14]

### 2.4. Transformation

Proses tokenisasi pada model IndoBERT menggunakan tokenisasi *WordPiece tokenizer*, di mana *tokenizer* yang dilatih sebelumnya lebih efektif untuk memproses kata-kata di luar kosakata yang dilatih. *Tokenizer* ini dapat menangani bentuk kata yang kurang terstruktur, dan bahasa gaul.[15] Model XGBoost menerapkan pendekatan statistik TF-IDF yang menilai urgensi suatu kata [16]. Tokenisasi oleh XGBoost dimulai dengan memisahkan teks berdasarkan *space* dan numerik, dan teks ditranslasikan menjadi *token* [17]. Pada proses ini diperlukan *data splitting* yang merupakan *dataset* akhir yang sudah dibersihkan dan diberi label, yang akan dipisah menjadi *training data* dan *testing data*, dengan rasio 80:20 atau *stratified sampling*.[18]

### 2.5. Data Mining

#### 2.5.1. Model IndoBERT

IndoBERT yang telah di *pre-trained* pada *dataset*, disesuaikan dengan *fine tuning* yang kemudian disimpan menjadi model baru.[19] Representasi metrik dari setiap data teks diekstrak menggunakan model IndoBERT, di mana data teks akan dikonversi menjadi *ID Token* dan akan dikalibrasi menggunakan *attention mask*. Fitur numerik akan diambil dari *embedding token*, yang menghasilkan dimensi metrik sebesar 768 untuk tiap baris data.[8]

#### 2.5.2. Model XGBoost

Fitur *embedding* dari IndoBERT akan diproses dengan algoritma XGBoost untuk menghasilkan probabilitas *multi-class*, dan efisiensi data dengan skala besar (*big-data)*. Proses dilakukan dengan pendekatan *traditional machine learning* dengan TF-IDF, yang tidak melalui tahap *pre-training* atau *fine-tuning*, namun dengan melatih model secara langsung dan mengujinya.[10]

### 2.6. Evaluation

Performa klasifikasi dari IndoBERT dan XGBoost dievaluasi berdasarkan pengujian terhadap *testing data*, yang berfokus ke *accuracy* (mengukur akurasi prediksi model dari *dataset*), *precision* (mengukur rasio antara data yang diprediksi benar positif dengan *dataset* yang diprediksi positif), *recall* (mengukur kemampuan model untuk menemukan kelas sentimen asli), dan *F1-score* (nilai rata-rata antara metrik *precision* dan *recall*).

## 3. HASIL DAN PEMBAHASAN

### 3.1. Exploration Data Analysis

Sebelum diberlakukan pembersihan data, dilakukan *Data Exploration* terhadap *dataset* yang berfungsi untuk menetapkan struktur data. Tahap ini dibagi menjadi dua metrik statistik utama, yaitu *character length*, dan *word count*

**Tabel 2.** Distribusi Statistik

| Statistic | Character Length | Word Count |
|---|---|---|
| count | 57362.000000 | 57362.000000 |
| mean | 66.341271 | 11.271713 |
| std | 84.596832 | 13.678566 |
| min | 0.000000 | 0.000000 |
| 25% | 21.000000 | 4.000000 |
| 50% | 42.000000 | 7.000000 |
| 75% | 82.000000 | 14.000000 |
| max | 2170.000000 | 351.000000 |

Tabel 2 membuktikan data memiliki rata-rata *length* sebesar 66,34 karakter dengan standar deviasi 84,60 karakter, yang menunjukkan adanya variasi sentimen masyarakat yang beragam. Hal yang sama dengan distribusi *word count* yang memiliki rata-rata 11,27 kata dengan standar deviasi 13,68 kata.

### 3.2. Noise Analysis

*Noise analysis* menggunakan ekspresi untuk mengidentifikasi sebaran data kotor dalam *dataset* yang menurunkan *performance* pada model saat tahap klasifikasi

**Tabel 3**. *Noise Analysis*

| No | Noise Type | Count |
|---|---|---|
| 1 | *Emoji* Only | 2450 |
| 2 | Contains *Mention* | 2196 |
| 3 | Contains URL | 49 |

Pada Tabel 3, menunjukkan adanya data kotor berupa data yang hanya berisi simbol *emoji* sebanyak 2,450 data, *mention* sebanyak 2,196 data, dan mengandung tautan sebanyak 49 data, yang merupakan salah satu fungsi dari *text preprocessing* yang berfungsi untuk menghapus data-data kotor pada *dataset*

**Algorithm Text Preprocessing**
Input: text, slang_dict
Output: cleaned_text

```
function preprocess_text(text, slang_dict)
  if text is not a String then
    return ""
  end if

  text ← lower_case(text)
  text ← regex_replace(r'http\S+|www\S+|https\S+|@\w+', '', text)
  text ← regex_replace(r'[^a-zA-Z\s]', '', text)
  text ← strip_and_collapse_whitespace(text)

  words ← split_by_space(text)
  normalized_words ← []

  for each word in words do
    if word exists in slang_dict then
      append slang_dict[word] to normalized_words
    else
      append word to normalized_words
    end if
  end for

  cleaned_text ← join_by_space(normalized_words)
  return cleaned_text
end function
```

### 3.3. Pseudo Labeling

*Pseudo labeling* dilakukan menggunakan model RoBERTa Bahasa Indonesia yang telah dilatih untuk analisis sentimen. Model ini dipakai untuk menghasilkan label awal beserta *confidence score* pada setiap

komentar. Model juga akan menghasilkan probabilitas setiap kelas dan *confidence score* yang akan digunakan pada tahap seleksi data. Lalu model akan menganalisis distribusi *confidence score* untuk mengevaluasi tingkat kelayakan model terhadap label yang dihasilkan.

**Tabel 4**. *Pseudo Labeling*

| No | Threshold | Total Data | Percentage (%) |
|---|---|---|---|
| 1 | 0.70 | 46829 | 82.33 |
| 2 | 0.75 | 44474 | 78.19 |
| 3 | 0.80 | 42023 | 73.88 |
| 4 | 0.85 | 38745 | 68.12 |
| 5 | 0.90 | 34898 | 61.36 |

Berdasarkan hasil pada tabel 4, nilai threshold sebesar 0,80 dipilih karena dinilai sebagai *trade-off* yang mampu mempertahankan sebagian besar data dan menyaring prediksi dengan *confidence* <80%, untuk meningkatkan kualitas *pseudo label* lebih baik sebelum digunakan pada proses *training*, dengan mempertahankan representasi data sebanyak 42.023 komentar (73.88%)

**Algorithm Pseudo Labeling & Confidence Filtering**
Input: df, Model RoBERTa, Threshold T = 0.80
Output: final_df

```
function run_pseudo_labeling(df, Model, T)
  predictions ← []
  batches ← divide_dataset_into_batches(df["comment"], batch_size=64)

  for each batch in batches do
    tokens ← tokenize_text(batch, max_length=128)
    logits ← forward_pass_RoBERTa(tokens)
    probabilities ← apply_softmax(logits, dimension=1)

    for each sample in batch do
      confidence ← max(probabilities[sample])
      pred_label ← argmax(probabilities[sample])

      if confidence ≥ T then
        append sample with pred_label and confidence to predictions
      end if
    end for
  end for

  final_df ← convert_to_dataframe(predictions)
  return final_df
end function
```

### 3.4. Data Splitting

*Dataset* dibagi menjadi *training data* dan *testing data* pengujian menggunakan rasio 80:20. Pembagian dilakukan dengan metode *stratified sampling* agar proporsi setiap kelas tetap terjaga.

**Algorithm Dataset Splitting**

Input: final_df["comment, final_df["label_encoded"], test_size = 0.20, SEED
Output: X_train, y_train, X_test, y_test
Training Data : 33.618
Testing Data : 8.405

```
function split_dataset(X, Y, test_size, SEED)
    shuffle the dataset randomly using SEED to ensure reproducibility
    split the dataset into training and testing sets based on test_size proportion
    apply stratification to maintain the identical ratio of classes from Y in both sets
    assign the partitioned subsets to X_train, X_test, y_train, and y_test
    return X_train, X_test, y_train, y_test
    end function
```

### 3.5. *Hybrid* Modeling

Evaluasi dilakukan menggunakan metrik *accuracy, precision, recall, dan F1-Score.* Dan juga menggunakan *confusion matrix* untuk menganalisis performa model pada masing-masing *class* sentimen setelah melakukan pelatihan pada 33.618 *training data*, dan 8.405 *testing data* berdasarkan pengujian tersebut, *overall performance* dari *hybrid model* sebagai berikut:[14].

**Tabel 5**. *Hybrid* Model Evaluation

| No | Metrics | Score |
|---|---|---|
| 1 | Accuracy | 0.952052 |
| 2 | Precision | 0.952409 |
| 3 | Recall | 0.952052 |
| 4 | F1-Score | 0.951812 |

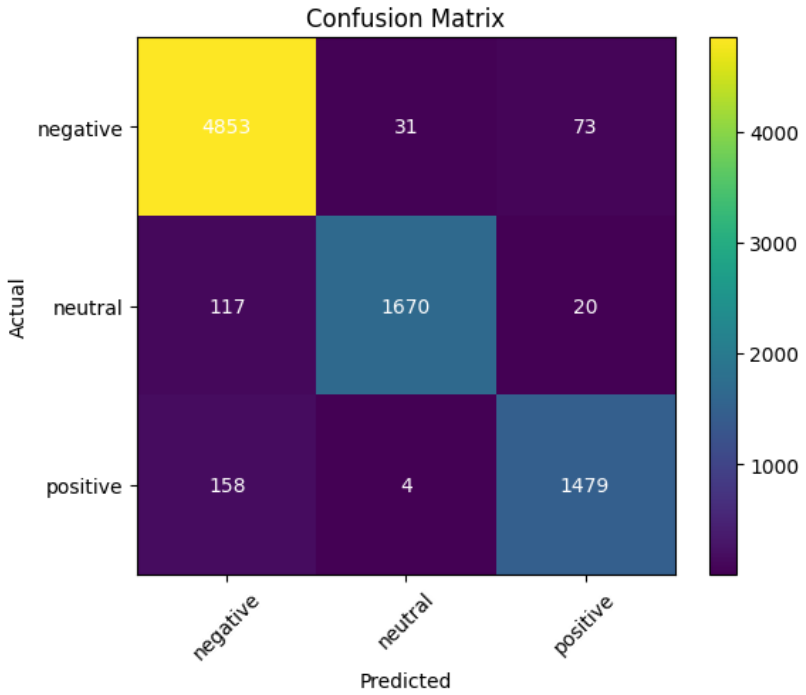
**Tabel 6.** *Dataframe Report*

| Sentiment | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| negative | 0.946373 | 0.979020 | 0.962419 | 4957.000000 |
| neutral | 0.979472 | 0.924184 | 0.951025 | 1807.000000 |
| positive | 0.940840 | 0.901280 | 0.920635 | 1641.000000 |
| accuracy | 0.952052 | 0.952052 | 0.952052 | 0.952052 |
| macro avg | 0.955562 | 0.934828 | 0.944693 | 8405.000000 |
| weighted avg | 0.952409 | 0.952052 | 0.951812 | 8405.000000 |

Model membuktikan deteksi pada *class* negatif unggul dengan metrik *recall* di angka 97,90% dan *F1-Score* 96,24%. Di mana model berhasil mengidentifikasi 4.853 data secara akurat dari 4.957 data. Performa *class* netral memiliki *F1-Score* 95,10%, dan *class* positif dengan *F1-Score* 92,06%.[4]

**Algorithm IndoBERT Feature Extraction**
Input: texts, IndoBERT_Tokenizer, IndoBERT_Model
Output: embeddings

```
function extract_embeddings(texts, IndoBERT_Tokenizer, IndoBERT_Model)
    embeddings ← []
    batches ← divide_into_batches(texts, batch_size=64)
    for each batch in batches do
        encoded_inputs ← IndoBERT_Tokenizer(batch, padding=True, truncation=True, max_length=128)
        outputs ← forward_pass_IndoBERT(encoded_inputs)
        cls_embedding ← outputs.last_hidden_state[:, 0, :]
        append cls_embedding to embeddings
    end for
    return stack_vertically(embeddings)
end function
```




**Gambar 2.** *Confusion Matrix*

### 3.6. Pembahasan

Berdasarkan Tabel 6 dan Gambar 2, hasil analisis menggunakan model *hybrid* IndoBERT dan XGBoost mampu mengatasi *data imbalance* pada *dataset* komentar TikTok MBG (Makan Bergizi Gratis). Walaupun *dataset* didominasi oleh *class* negatif dengan data sebanyak 4.957 sampel, dibandingkan *class* netral dengan 1.807 dan *class* positif dengan 1.641, klasifikasi tidak mengalami bias

Model IndoBERT membuktikan mampu mengekstrak data dengan bahasa yang kurang terstruktur dengan sangat baik, hasil menunjukan *F1-score* pada *weighted average* di 95,18%, dan algoritma XGBoost tidak perlu mengorbankan sensitivitas deteksi pada *class* minoritas.

## 4. KESIMPULAN

Penelitian ini berhasil mengimplementasikan dan mengevaluasi kinerja model hybrid IndoBERT dan XGBoost dalam menganalisis sentimen masyarakat terhadap program Makan Bergizi Gratis (MBG) pada media sosial TikTok. Dengan tahapan penelitian berupa pengumpulan data, exploration data analysis , text preprocessing, pseudo labeling menggunakan model RoBERTa, ekstraksi fitur dengan IndoBERT, dan klasifikasi menggunakan XGBoost terhadap 42.023 data komentar setelah *confidence filtering*. Model hybrid ini mencatat angka *accuracy* sebesar 95,21% dan *F1-score* di 95,18% saat menguji 8.405 sampel testing.[22]

Hasil analisis pembagian per class membuktikan dominasi performa pada class sentimen negatif, tanpa mengalami bias klasifikasi pada kelompok sentimen netral dan positif.[21] Membuktikan bahwa algoritma XGBoost dan model IndoBERT mampu mengekstrak dan mengklasifikasikan fitur teks bahasa Indonesia yang kurang terstruktur di media sosial.[7]

## PERNYATAAN KETERSEDIAAN DATA

Data yang kami gunakan sebagai referensi dalam penulisan *paper* ini terdapat pada Daftar Pustaka, kami bersedia untuk menyerahkan seluruh kumpulan data, kode, atau materi pendukung lainnya pada repositori yang dapat diakses publik.

Kami bertanggung jawab atas keakuratan, integritas, dan dokumentasi yang tepat dari data yang digunakan dalam penelitian ini.

## UCAPAN TERIMA KASIH DAN PENGUNGKAPAN PENGGUNAAN AI

Kami berterima kasih untuk seluruh pihak yang terlibat dalam pendanaan dan penyusunan *paper* ini, Kami juga mengucapkan terima kasih kepada pihak kampus yang telah memberikan fasilitas dalam publikasi *paper* ini.

Dalam penulisan *paper* ini, tidak ada alat kecerdasan buatan (AI) atau teknologi yang dibantu AI yang digunakan dalam penyusunan *paper* ini.

## KONFLIK KEPENTINGAN

Para penulis menyatakan bahwa tidak ada konflik kepentingan terkait publikasi makalah ini.

## DAFTAR PUSTAKA

[1] B. Wilie, K. Vincentio, G. I. Winata, A. Wibowo, A. Lim, S. Chiong, and A. Purwarianti, “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding,” pp. 843–857, 2020.

[2] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[3] A. F. Al Farizi and Y. Sibaroni, “Implementation of BiLSTM and IndoBERT for Sentiment Analysis of TikTok Reviews,” JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), vol. 10, no. 1, pp. 96–106, 2025.

[4] B. P. Aji, C. Sri, and K. Aditya, “Klasifikasi Sentimen Ulasan Produk pada Platform E-Commerce di Indonesia dengan Menggunakan Model Pre-Trained IndoBERT,” Technology and Science (BITS), vol. 6, no. 4, 2025.

[5] F. M. Apriansyah, T. I. Ramadhan, C. R. Hidayat, and A. K. Wijaya, “Perbandingan IndoBERT dan IndoRoBERTa Untuk Analisis Sentimen Pada Film Dokumenter Dirty Vote,” Jurnal Informatika: Jurnal Pengembangan IT, vol. 10, no. 3, pp. 593–605, 2025.

[6] A. Y. Pratama, G. A. Sanjaya, N. K. Lubis, M. R. Aditya, and Y. Yennimar, “Analisis Sentimen Publik Terkait Danantara Menggunakan Algoritma IndoBERT pada Platform Media Sosial,” METIK Jurnal, vol. 9, no. 1, pp. 92–100, 2025.

[7] G. F. Situmorang and R. Purba, “Deteksi Potensi Depresi dari Unggahan Media Sosial X Menggunakan IndoBERT,” Building of Informatics, Technology and Science (BITS), vol. 6, no. 2, pp. 649–661, Sep. 2024.

[8] M. I. Rizaldi, D. R. Chandranegara, and D. R. Akbi, “Comparison Of Machine Learning Techniques For Classification Of Distributed Denial Of Service Attacks Based On Feature Engineering In Sdn-Based Networks,” JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), vol. 9, no. 3, pp. 1180–1197, 2024.

[9] I. A. Latifah, F. A. Rafrastara, J. Bintoro, W. Ghozi, and W. M. Osman, “Comparative Analysis of Feature Selection Methods with XGBoost for Malware Detection on the Drebin Dataset,” Jurnal Sisfokom (Sistem Informasi dan Komputer), vol. 13, no. 3, pp. 403–409, 2024.

[10] S. Rosyada, F. A. Rafrastara, A. Ramadhani, W. Ghozi, and W. Yassin, “Enhancing XGBoost Performance in Malware Detection through Chi-Squared Feature Selection,” Jurnal Sisfokom (Sistem Informasi dan Komputer), vol. 13, no. 3, pp. 396–402, 2024.

[11] S. B. Bulkisah, R. Astuti, and A. Bahtiar, “Implementasi Data Mining Algoritma Decision Tree Untuk Klasifikasi Status Gizi Balita Di Kecamatan Ciledug,” Jurnal Ilmiah Informatika Komputer, vol. 29, no. 1, pp. 1–12, 2024.

[12] A. S. Aribowo, H. Basiron, N. F. Abd Yusof, and S. Khomsah, “Cross-domain sentiment analysis model on Indonesian YouTube comment.,” International Journal of Advances in Intelligent Informatics, vol. 7, no. 1, 2021.

[13] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, “Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19),” J. Media Inform. Budidarma, vol. 5, no. 2, pp. 406, 2021.

[14] L. R. Krosuri and R. S. Aravapalli, “Novel heuristic-based hybrid ResNeXt with recurrent neural network to handle multi class classification of sentiment analysis,” Mach Learn Sci Technol, vol. 4, no. 1, p. 015033, 2023.

[15] F. R. Andhika, W. Witanti, and P. N. Sabrina, “Analisis Sentimen Menggunakan Metode IndoBERT Pada Ulasan Aplikasi Zoom Menggunakan Fitur Ekstrasi GloVe,” METIK Jurnal, vol. 9, no. 2, pp. 439–448, 2025.

[16] Y. Yulistiani and S. Styawati, “Analisis Sentimen Terhadap Calon Presiden Indonesia 2024 dengan Metode Extreme Gradient Boosting (XGBoost),” Jurnal Informatika: Jurnal Pengembangan IT, vol. 9, no. 3, pp. 322–328, 2024.

[17] A. Rácz, D. Bajusz, and K. Héberger, “Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification,” Molecules, vol. 26, no. 4, p. 1111, 2021.

[18] 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. 1706–1714, 2025.

[19] A. Saputri and D. Utari, "Efektivitas Program Makan Bergizi Gratis (MBG) Terhadap Status Gizi, Kesiapan Belajar, Dan Luaran Pendidikan Peserta Didik: Tinjauan Sistematis," vol. 14, no. 1, pp. 35–48, May 2026.

[20] R. A. Hanasi, M. R. Tuna, D. P. Burhan, M. O. Sila, A. Ladjulu, and S. N. Abuba, "Analisis Efektivitas Program Makan Bergizi Gratis (MBG) di Indonesia sebagai Instrumen Kebijakan Pemerintah," *Jurnal Pendidikan Tambusai*, vol. 9, no. 3, pp. 38022–38028, 2025.

[21] N. Putri, D. Kurniawan, and K. Tania, "Deteksi Komentar dan Analisis Sentimen Promosi Judi Online pada Youtube Menggunakan IndoBERT dan XGBoost," *Building of Informatics, Technology and Science (BITS)*, vol. 7, no. 3, pp. 1592–1600, 2025.

[22] J. A. Lukmando and P. Theresia, "Analisis Sentimen Program Makan Siang Gratis Menggunakan Model IndoBERT," *Jurnal Al-Azhar Indonesia Seri Sains dan Teknologi*, vol. 8, no. 2, pp. 820–829, 2025.