

ABSTRAK

ANALISIS SENTIMEN KOMENTAR YOUTUBE TERHADAP PROGRAM MAKAN BERGIZI GRATIS MENGGUNAKAN ALGORITMA NAIVE BAYES

Nama : Laksamana Rhazes Syawal
NPM : 2255201137
Pembimbing : Dr. Sastya Hendri Wibowo, S.Kom., M.Kom.

Program Makan Bergizi Gratis (MBG) merupakan salah satu kebijakan strategis pemerintah Indonesia yang bertujuan untuk meningkatkan kualitas gizi anak usia sekolah di seluruh wilayah Indonesia. Program ini memicu beragam respons masyarakat, baik berupa dukungan maupun kritik, yang secara aktif diungkapkan melalui komentar pada platform YouTube. Penelitian ini bertujuan untuk menganalisis kecenderungan sentimen masyarakat terhadap Program MBG berdasarkan komentar YouTube menggunakan algoritma *Naive Bayes* sebagai metode klasifikasi, dengan dukungan teknik *Easy Data Augmentation* (EDA) untuk mengatasi ketidakseimbangan data serta pelabelan otomatis menggunakan model IndoBERT (Indonesian RoBERTa) sebagai pendekatan *pseudo-labelling*. Data penelitian dikumpulkan melalui teknik scraping menggunakan YouTube DATA API v3 dengan 18 variasi kata kunci yang relevan sehingga menghasilkan 10.073 komentar unik. Setelah melalui proses seleksi dan prapemrosesan teks yang meliputi *cleaning*, *case folding*, *normalizing*, *tokenizing*, *stopword removal*, dan *stemming* menggunakan library Sastrawi, diperoleh 9.589 data yang siap dianalisis. Selanjutnya, data diberi label secara otomatis menggunakan model IndoBERT (w11wo/indonesian-roberta-base-sentiment-classifier) dengan distribusi 5.039 komentar negatif (52,5%), 2.517 komentar netral (26,3%), dan 2.032 komentar positif (21,2%). Untuk mengatasi ketidakseimbangan distribusi kelas, diterapkan teknik *Easy Data Augmentation* (EDA) menggunakan strategi *Random Word Swap* ($\text{aug_p}=0.2$), sehingga jumlah data pada setiap kelas menjadi seimbang, yaitu masing-masing 5.039 data. Proses ekstraksi fitur dilakukan menggunakan metode TF-IDF dengan konfigurasi N-Gram (1,2), sedangkan pemilihan hyperparameter menggunakan GridSearchCV dengan validasi silang 5-fold. Hasil pengujian menunjukkan bahwa model *Naive Bayes* dengan nilai α sebesar 0,1 mampu mencapai *accuracy* keseluruhan sebesar 84,06%, dengan nilai *F1-Score* masing-masing sebesar 0,81 untuk kelas negatif, 0,83 untuk kelas netral, dan 0,88 untuk kelas positif. Temuan ini menunjukkan bahwa kombinasi algoritma *Naive Bayes*, teknik EDA, dan pelabelan otomatis menggunakan IndoBERT efektif dalam menganalisis sentimen masyarakat terhadap Program Makan Bergizi Gratis berdasarkan komentar pada platform YouTube.

Kata Kunci: Analisis Sentimen, Program Makan Bergizi Gratis, YouTube, *Naive Bayes*, TF-IDF, *Easy Data Augmentation* (EDA), IndoBERT, *Pseudo-Labelling*, CRISP-DM.

ABSTRACT

SENTIMENT ANALYSIS OF YOUTUBE COMMENTS ON THE FREE NUTRITIOUS MEAL PROGRAM USING THE NAIVE BAYES ALGORITHM

NAME : Laksamana Rhazes Syawal

STUDENT ID : 2255201137

SUPERVISOR : Dr. Sastya Hendri Wibowo, S.Kom., M.Kom.

The Free Nutritious Meal Program (Program Makan Bergizi Gratis or MBG) is one of the Indonesian government's strategic policies aimed at improving the nutritional quality of school-age children across Indonesia. The program has generated diverse public responses, including both support and criticism, which are actively expressed through comments on the YouTube platform. This study aimed to analyze public sentiment toward the MBG Program based on YouTube comments using the *Naive Bayes* algorithm as a classification method, supported by *Easy Data Augmentation* (EDA) to address data imbalance and automatic labeling using the IndoBERT (Indonesian RoBERTa) model through a *pseudo-labelling* approach. The research data were collected through scraping using the YouTube Data API v3 with 18 variations of relevant keywords, resulting in 10,073 unique comments. After selection and text preprocessing, including *cleaning*, *case folding*, *normalization*, *tokenization*, *stopword removal*, and *stemming* using the Sastrawi library, 9,589 data instances were obtained for analysis. The data were then automatically labeled using the IndoBERT model (*w11wo/indonesian-roberta-base-sentiment-classifier*), resulting in 5,039 negative comments (52.5%), 2,517 neutral comments (26.3%), and 2,032 positive comments (21.2%). To address the class imbalance, *Easy Data Augmentation* (EDA) was applied using the *Random Word Swap* strategy ($\text{aug_p} = 0.2$), resulting in a balanced dataset of 5,039 instances for each class. Feature extraction was performed using the Term Frequency-Inverse Document Frequency (TF-IDF) method with an N-Gram configuration of (1,2), while hyperparameter selection was conducted using GridSearchCV with 5-fold cross-validation. The evaluation results showed that the *Naive Bayes* model with an alpha value of 0.1 achieved an overall accuracy of 84.06%, with F1-scores of 0.81 for the negative class, 0.83 for the neutral class, and 0.88 for the positive class. These findings indicate that the combination of the *Naive Bayes* algorithm, EDA technique, and automatic labeling using IndoBERT is effective for analyzing public sentiment toward the Free Nutritious Meal Program based on comments on the YouTube platform.

Keywords: Sentiment Analysis, Free Nutritious Meal Program, YouTube, *Naive Bayes*, TF-IDF, *Easy Data Augmentation* (EDA), IndoBERT, *Pseudo-Labelling*, CRISP-DM

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi dan komunikasi telah mengubah media sosial menjadi sarana utama dalam memperoleh dan menyebarkan informasi. Media sosial kini telah menjadi alat komunikasi yang banyak digunakan masyarakat untuk mengungkapkan opini, dukungan dan kritik tentang berbagai isu politik dan kebijakan pemerintah. Platform YouTube merupakan salah satu media sosial berbasis video yang banyak dimanfaatkan masyarakat untuk memperoleh informasi. Platform ini memungkinkan pengguna mengunggah video, sementara pengguna lain dapat memberikan komentar, serta menekan tombol *like* atau *dislike* pada konten tersebut. Setiap kebijakan pemerintah yang diumumkan atau dilaksanakan berpotensi menimbulkan berbagai tanggapan dari masyarakat melalui komentar dan unggahan di platform tersebut (Pratama *et al.*, 2023).

Program Makan Bergizi Gratis (MBG) merupakan salah satu kebijakan pemerintah yang memiliki dampak yang signifikan terhadap perekonomian dan kehidupan masyarakat. Program Makan Bergizi Gratis (MBG) merupakan program pemerintah yang sering menjadi sorotan publik, karena berdampak luas pada berbagai aspek kehidupan. Program semacam itu dapat menyebabkan perubahan harga kebutuhan pokok, serta memengaruhi perekonomian secara keseluruhan. Akibatnya, program ini telah mendapat perhatian luas dari publik dan banyak dibahas di media sosial. Sebagian masyarakat memberikan dukungan dan apresiasi, sementara sebagian lainnya menyampaikan kritik terkait kualitas,

distribusi maupun pelaksanaan program di lapangan (Rahmatullah *et al.*, 2025; Wahyuni *et al.*, 2025).

Selama program ini berlangsung tidak akan terlepas dari berbagai tanggapan publik di media sosial. Perhatian publik meningkat seiring munculnya isu keracunan makanan dan dugaan kasus korupsi yang dilakukan mantan petinggi Badan Gizi Nasional (BGN) yang menjadi topik hangat di YouTube. Kondisi ini menunjukkan bahwa media sosial telah menjadi sarana utama bagi masyarakat untuk menyampaikan pendapat terkait program pemerintah ini.

Komentar pengguna pada platform YouTube merupakan data tekstual yang bersifat tidak terstruktur dan memiliki karakteristik bahasa yang beragam, cenderung informal serta sering mengandung singkatan, jargon dan ungkapan emosional. Banyaknya komentar yang muncul dalam waktu relatif singkat menyebabkan proses analisis opini secara manual kurang efektif dan efisien. Oleh karena itu, dibutuhkan metode yang efektif untuk mengelola dan menganalisis data teks dalam skala besar secara terstruktur (Rahmatullah *et al.*, 2025).

Analisis sentimen adalah bagian dari penambangan teks (*text mining*) dan pemrosesan bahasa alami (*natural language processing*) yang bertujuan untuk mengidentifikasi dan mengklasifikasikan opini masyarakat terhadap suatu objek ke dalam kategori tertentu, seperti positif, negatif dan netral. Pendekatan ini memungkinkan data teks yang tidak terstruktur diolah menjadi informasi yang lebih sistematis dan terstruktur (Nurwanda *et al.*, 2024).

Algoritma *Naive Bayes* dipilih karena dipilih sebagai metode klasifikasi dalam penelitian ini karena telah terbukti memiliki performa yang baik dalam

klasifikasi teks berbahasa Indonesia, dengan keunggulan berupa proses komputasi yang sederhana, cepat dan efektif dalam mengolah data berukuran besar (Prastiawan & Yuniarto, 2025; Rahmatullah *et al.*, 2025). Metode ini dikenal memiliki proses komputasi yang sederhana, cepat dan efektif dalam mengolah data teks dengan jumlah besar. Untuk meningkatkan kualitas pelabelan data, penelitian ini menggunakan model IndoBERT (Indonesian RoBERTa), yaitu model pra-latih berbasis arsitektur Transformer yang telah di-fine-tuning khusus untuk klasifikasi sentimen Bahasa Indonesia, sebagai pendekatan *pseudo-labelling* yang menggantikan pelabelan manual yang rentan terhadap subjektivitas. Selain itu, untuk mengatasi permasalahan ketidakseimbangan distribusi kelas (*class imbalance*) yang umum terjadi pada data sentimen media sosial, diterapkan teknik *Easy Data Augmentation* (EDA) yang mampu menciptakan data sintetis untuk menyeimbangkan jumlah data antar kelas tanpa mengubah makna semantik teks asli (Wei & Zou, 2019).

Dengan menerapkan kombinasi metode tersebut, hasil penelitian ini diharapkan dapat memberikan kontribusi dalam memahami kecenderungan opini publik terhadap Program Makan Bergizi Gratis, sekaligus mendemonstrasikan efektivitas pendekatan *hybrid* yang mengintegrasikan model *deep learning* (IndoBERT) sebagai pelabel dengan algoritma *Naive Bayes* sebagai klasifikasi dalam konteks analisis sentimen berbahasa Indonesia.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut :

1. Bagaimana kecenderungan dan distribusi sentimen masyarakat terhadap Program Makan Bergizi Gratis (MBG) berdasarkan komentar yang diperoleh dari platform YouTube?
2. Bagaimana penerapan algoritma *Naive Bayes* dengan teknik *Easy Data Augmentation* (EDA) dan pelabelan IndoBERT (*pseudo-labelling*) dalam mengklasifikasikan sentimen komentar YouTube terhadap Program Makan Bergizi Gratis (MBG)?
3. Bagaimana kinerja algoritma *Naive Bayes* dalam mengklasifikasikan sentimen komentar YouTube terhadap Program Makan Bergizi Gratis (MBG) berdasarkan metrik evaluasi yang digunakan?

1.3 Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut :

1. Menganalisis kecenderungan sentimen masyarakat terhadap Program Makan Bergizi Gratis berdasarkan komentar yang diperoleh dari platform YouTube.
2. Menerapkan algoritma *Naive Bayes* dengan teknik *Easy Data Augmentation* (EDA) dan pelabelan IndoBERT (*pseudo-labelling*) dalam proses klasifikasi sentimen komentar pengguna.

3. Mengevaluasi tingkat akurasi dan efektivitas model klasifikasi berdasarkan metrik evaluasi yang digunakan, yaitu *accuracy*, *precision*, *recall* dan *F1-Score*.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan beberapa manfaat sebagai berikut :

1. Bagi peneliti, penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan ilmu pengetahuan di bidang analisis sentimen, *machine learning*, *text mining*, khususnya dalam penerapan algoritma *Naive Bayes* yang dikombinasikan dengan teknik *Text Augmentation* (EDA) dan pelabelan otomatis menggunakan model IndoBERT pada data teks berbahasa Indonesia.
2. Bagi akademis, penelitian ini dapat menjadi referensi bagi mahasiswa maupun peneliti lain yang tertarik pada topik analisis sentimen dan klasifikasi teks. Penelitian yang dilakukan dapat mendukung pengembangan studi lanjutan dengan metode, algoritma atau pendekatan yang berbeda.
3. Bagi pemerintah dan pengelola program, penelitian dilakukan berfokus pada analisis kecenderungan respons publik terhadap Program Makan Bergizi Gratis berdasarkan data yang dianalisis secara sistematis. Informasi tersebut dapat dimanfaatkan untuk menunjang proses penentuan langkah strategis lebih lanjut dalam pelaksanaan kebijakan publik.

1.5 Ruang Lingkup Penelitian

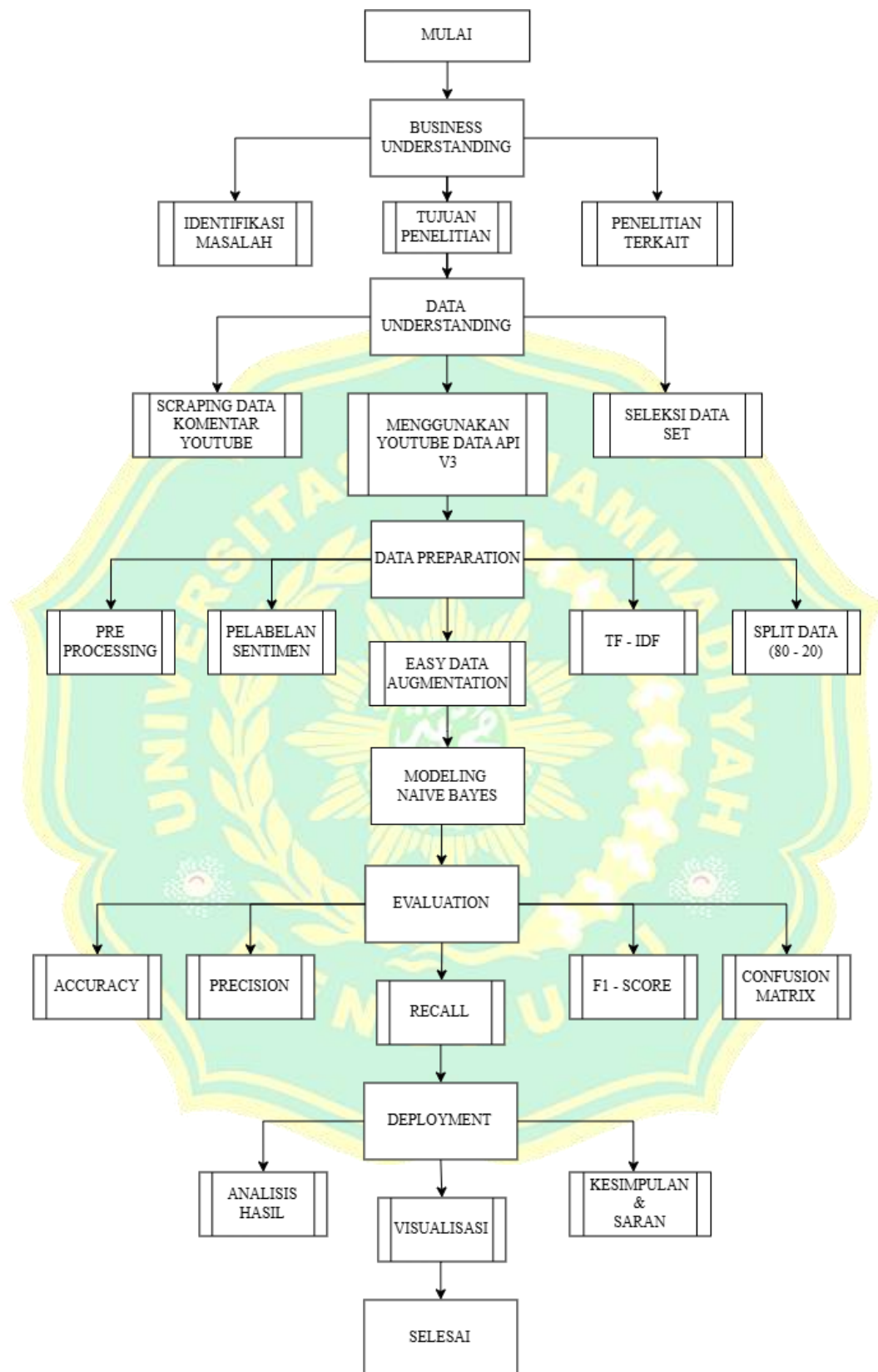
Ruang lingkup diperlukan agar penelitian tetap fokus dan tidak melebar dari tujuan yang telah ditetapkan. Pembatasan ini bertujuan untuk memperjelas batas analisis yang dilakukan sehingga penelitian ini dapat dilaksanakan secara terarah dan sistematis. Ruang lingkup penelitian ini meliputi :

1. Data penelitian ini berupa komentar pengguna YouTube seputar Program Makan Bergizi Gratis yang dikumpulkan melalui YouTube DATA API v3.
2. Data yang dianalisis dibatasi pada komentar berbahasa Indonesia.
3. Penelitian ini berfokus pada analisis sentimen komentar berdasarkan tiga kategori, yaitu sentimen positif, negatif dan netral.
4. Pelabelan sentimen dilakukan secara otomatis menggunakan model IndoBERT (*pseudo-labelling*) sebagai pengganti pelabelan manual.
5. Diterapkan teknik *Easy Data Augmentation* (EDA) dengan strategi *Random Word Swap* untuk mengatasi ketidakseimbangan distribusi kelas.
6. Analisis difokuskan pada data berbasis teks tanpa mempertimbangkan unsur lain seperti profil pengguna, jumlah *like*, jumlah *subscriber* atau faktor eksternal lainnya.
7. Evaluasi performa model menggunakan metrik *accuracy*, *precision*, *recall* dan *F1-Score*.

Pembatasan tersebut dilakukan agar penelitian tetap terarah pada analisis sentimen berbasis teks dan tidak melebar ke analisis multimodal atau metode klasifikasi lainnya.

1.6 Framework Penelitian

Framework penelitian ini dirancang berdasarkan metodologi CRISP-DM. Data berupa komentar masyarakat pada platform YouTube dikumpulkan melalui teknik *scraping* menggunakan YouTube DATA API v3. Data yang berhasil diperoleh kemudian diproses melalui tahapan *preprocessing* (*cleaning, case folding, normalizing, tokenizing, stopword removal* dan *stemming*), pelabelan sentimen menggunakan model IndoBERT, *Easy Data Augmentation* (EDA) untuk penyeimbangan data, ekstraksi fitur menggunakan TF-IDF, serta pembagian data set menjadi data latih dan data uji. Selanjutnya, model klasifikasi dibangun menggunakan algoritma Multinomial *Naive Bayes* dengan optimasi hyperparameter melalui GridSearchCV. Kinerja model dievaluasi menggunakan *Confusion Matrix* dengan metrik *accuracy, precision, recall* dan *F1-Score*. Framework penelitian ini dapat dilihat pada gambar dibawah ini.



GAMBAR 1.1 Framework Penelitian