

ABSTRAK

Media sosial, khususnya YouTube, telah menjadi sumber opini publik yang masif terhadap berbagai kebijakan pemerintah, termasuk kebijakan Makan Bergizi Gratis (MBG). Penelitian ini bertujuan untuk menerapkan metode *ensemble learning* dengan pendekatan *confidence-based soft voting* dalam klasifikasi sentimen opini publik pada komentar YouTube terhadap kebijakan MBG, serta membandingkan performanya dengan model *Logistic Regression* dan IndoBERT secara individual. Data penelitian berupa 963 komentar YouTube yang dikumpulkan menggunakan YouTube Data API v3, kemudian dilabeli secara manual melalui pendekatan *collaborative labeling* dengan melibatkan dua anotator.

Komentar diklasifikasikan ke dalam tiga kategori sentimen: negatif (369 data), positif (278 data), dan netral (316 data). Proses *preprocessing* teks mencakup tujuh tahapan, yaitu *case folding*, *text cleaning*, tokenisasi, normalisasi teks, *negation handling*, *stopword removal*, dan *stemming*. Representasi fitur untuk *Logistic Regression* menggunakan TF-IDF dengan *max_features*=5.000, sedangkan IndoBERT menggunakan *fine-tuning* pada model *indobenchmark/indobert-base-p1*. Metode *confidence-based soft voting* menggabungkan probabilitas prediksi kedua model menggunakan bobot dinamis berbasis nilai *confidence* masing-masing model.

Hasil pengujian pada 289 data uji menunjukkan bahwa *Logistic Regression* memperoleh akurasi 72%, IndoBERT 83%, dan *Ensemble Confidence-Based Soft Voting* 84% dengan *weighted F1-score* sebesar 0,84. Penelitian ini membuktikan bahwa kombinasi model klasik dan *transformer* dalam sistem *ensemble* mampu meningkatkan performa klasifikasi sentimen berbahasa Indonesia.

Kata kunci: analisis sentimen, *ensemble learning*, *soft voting*, IndoBERT, *Logistic Regression*, TF-IDF, Makan Bergizi Gratis, YouTube

ABSTRACT

Social media, particularly YouTube, has become a massive source of public opinion regarding various government policies, including the Makan Bergizi Gratis (MBG) program. This study aims to implement an ensemble learning method with a confidence-based soft voting approach for sentiment classification of public opinions in YouTube comments related to the MBG policy, and to compare its performance against individual Logistic Regression and IndoBERT models. The research data comprised 963 YouTube comments collected using the YouTube Data API v3, which were manually labeled through a collaborative labeling approach involving two annotators.

Comments were classified into three sentiment categories: negative (369 data), positive (278 data), and neutral (316 data). Text preprocessing involved seven stages: case folding, text cleaning, tokenization, text normalization, negation handling, stopwords removal, and stemming. Feature representation for Logistic Regression employed TF-IDF with `max_features=5,000`, while IndoBERT utilized fine-tuning on the `indobenchmark/indobert-base-p1` model. The confidence-based soft voting method combines prediction probabilities from both models using dynamic weights based on each model's confidence value.

Test results on 289 test samples showed that Logistic Regression achieved 72% accuracy, IndoBERT 83%, and the Ensemble Confidence-Based Soft Voting achieved 84% accuracy with a weighted F1-score of 0.84. This research demonstrates that combining classical and transformer-based models in an ensemble system can improve Indonesian sentiment classification performance.

Keywords: sentiment analysis, ensemble learning, soft voting, IndoBERT, Logistic Regression, TF-IDF, Makan Bergizi Gratis, YouTube