ABSTRAK

Proyek Kereta Cepat Jakarta-Bandung (Whoosh) dengan nilai investasi yang masif memicu perdebatan fiskal yang dinamis di ruang publik digital, khususnya terkait beban utang negara. Penelitian ini bertujuan menganalisis karakteristik sentimen masyarakat terhadap isu beban utang tersebut secara komprehensif melalui *Multi-platform* media sosial (X, YouTube, dan TikTok). Pendekatan yang digunakan adalah model *Hybrid Natural Language Processing (NLP)* berbasis IndoBERT dan Support Vector Machine (SVM) yang dikembangkan di bawah kerangka kerja CRISP-DM. Sebanyak 6.766 baris data tekstual bersih dikumpulkan selama periode Juli 2025 hingga Januari 2026. Fitur semantik teks diekstraksi menggunakan *CLS Embeddings* (768 dimensi) dari model *Pre-trained indobenchmark/indobert-base-p1*. Guna mengatasi masalah ketidakseimbangan kelas secara objektif, teknik undersampling dan SMOTE dievaluasi dengan kepatuhan pipeline yang ketat hanya pada ruang data latih untuk menghindari bias *data leakage*. Hasil eksperimen menunjukkan bahwa model SVM dengan Kernel Polynomial (POLY) yang dilatih pada dataset original (split 90:10) keluar sebagai konfigurasi terbaik. Model ini mencatatkan akurasi test sebesar 87.44% dan Macro F1 sebesar 0.7830, dengan rata-rata akurasi *5-Fold Stratified Cross Validation* sebesar 86.83% serta gap generalisasi (Train-CV) yang terkendali sebesar 4.55%. Dari sisi substantif, opini publik didominasi secara mutlak oleh sentimen negatif (70.0%), disusul netral (17.47%), dan positif (12.53%). Analisis tekstual mengindikasikan bahwa kecemasan utama masyarakat berakar pada beban fiskal APBN, isu transparansi tata kelola proyek, serta ketergantungan geopolitik terhadap pendanaan Tiongkok.

Kata Kunci: Analisis Sentimen, Kereta Cepat Whoosh, IndoBERT, Support Vector Machine, CRISP-DM, *Multi-platform* Media Sosial, Ketidakseimbangan Kelas.

ABSTRACT

The Jakarta-Bandung High-Speed Rail (Whoosh) project, with its massive investment value, has triggered a dynamic fiscal debate in the digital public sphere, particularly regarding the national debt burden. This study aims to comprehensively analyze the characteristics of public sentiment toward the debt burden issue across multiple social media platforms (X, YouTube, and TikTok). The approach used is a Hybrid Natural Language Processing (NLP) model based on IndoBERT and Support Vector Machine (SVM) developed under the CRISP-DM framework. A total of 6,766 clean textual data rows were collected during the period of July 2025 to January 2026. Semantic features of the text were extracted using CLS Embeddings (768 dimensions) from the Pre-trained indobenchmark/indobert-base-p1 model. To address the class imbalance issue objectively, undersampling and SMOTE techniques were evaluated with strict pipeline compliance only on the training set data space to avoid data leakage bias. The experimental results show that the SVM model with a Polynomial (POLY) Kernel trained on the original dataset (90:10 split) emerged as the best configuration. This model recorded a test accuracy of 87.44% and a Macro F1 of 0.7830, with an average 5-Fold Stratified Cross Validation accuracy of 86.83% and a controlled generalization gap (Train-CV) of 4.55%. Substantively, public opinion is absolutely dominated by negative sentiment (70.0%), followed by neutral (17.47%), and positive (12.53%). Textual analysis indicates that the main public anxieties are rooted in the state budget (APBN) fiscal burden, project governance transparency issues, and geopolitical dependence on Chinese funding.

Keywords: *Sentiment Analysis, Whoosh High-Speed Train, IndoBERT, Support Vector Machine, CRISP-DM, Multi-platform Social Media, Class Imbalance*.