ABSTRAK

Darmawan, Ardhi Amin. 2026. Analisis Sentimen Ulasan *Sound Horeg* Menggunakan Algoritma *Random Forest.* Tugas Akhir. Program Sarjana. Program Studi Teknik Informatika. Universitas Muhammadiyah Jember.
Pembimbing: Deni Arifianto, M.Kom.; Dr. Ilham Saifudin, S.Pd., M.Si.

Sound Horeg telah menjadi fenomena yang banyak dibahas di platform media sosial X, memicu apresiasi sekaligus keluhan di kalangan pengguna internet Indonesia namun, opini publik mengenai isu ini sebagian besar belum terdokumentasi dalam bentuk yang terukur. Penelitian ini bertujuan untuk mengklasifikasikan sentimen publik terhadap "*soundHoreg*" ke dalam kategori positif, negatif, dan netral menggunakan algoritma *Random Forest*. Data dikumpulkan melalui *Tweet Harvest* dengan kata kunci "*#Horeg*" selama periode 1 Januari hingga 15 Agustus 2025, yang menghasilkan 7.007 unggahan mentah. Pelabelan data dilakukan secara kombinasi (*semi-automated labeling*) menggunakan leksikon *TextBlob* dengan *threshold* polaritas $P > 0{,}05$ (positif), $P < -0{,}05$ (negatif), dan $-0{,}05 \leq P \leq 0{,}05$ (netral), yang divalidasi oleh pakar NLP (Hardian Oktavianto, S.Si., M.Kom.) terhadap 700 sampel acak berstrata (10% data) dengan tingkat kesepakatan *Cohen's Kappa* $\kappa = 0{,}84$ (*almost perfect agreement*). Sebaran label akhir terdiri dari 6.457 ulasan netral, 477 ulasan positif, dan 73 ulasan negatif. Teks melalui lima tahap pra-pemrosesan (*cleaning*, *tokenizing*, *stopword removal*, normalisasi, dan *stemming*) dan diekstraksi menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF). Guna menangani ketimpangan data, teknik *Synthetic Minority Over-sampling Technique* (SMOTE) diterapkan hingga tiap kelas berjumlah 6.457 sampel. Evaluasi dilakukan menggunakan 5-Fold *Cross-Validation* pada rasio 70:30, 80:20, dan 90:10. Hasil penelitian menunjukkan bahwa rasio 80:20 menghasilkan performa paling optimal dengan tingkat akurasi mencapai 99,98% serta nilai presisi, *recall*, dan *F1-score* berkisar antara 0,99 hingga 1,00 pada seluruh kelas sentimen. Analisis kualitatif terhadap pembobotan fitur TF-IDF mengonfirmasi bahwa keputusan klasifikasi didorong oleh kata kunci berdampak fisik pada kelas negatif (parah, kesal, marah, bising) dan kata apresiasi pada kelas positif (mantap, senang, suka, hiburan). Penelitian ini memberikan gambaran komprehensif dan objektif berbasis data mengenai persepsi publik terhadap *sound Horeg*.

Kata Kunci: *Sentiment Analysis*, *Random Forest*, TF-IDF, SMOTE, *Sound Horeg*

ABSTRACT

Darmawan, Ardhi Amin. 2026. *Sentiment Analysis of Sound Horeg Reviews Using the Random Forest Algorithm*. Final Project. Undergraduate Program. Informatics Engineering Study Program. Universitas Muhammadiyah Jember.

Supervisors: Deni Arifianto, M.Kom.; Dr. Ilham Saifudin, S.Pd., M.Si.

Sound Horeg has become a widely discussed phenomenon on the social media platform X, sparking both appreciation and complaints among Indonesian internet users. However, public opinion on this issue has largely not been documented in a measurable form. This study aims to classify public sentiment toward "sound Horeg" into positive, negative, and neutral categories using the Random Forest algorithm. Data were collected using Tweet Harvest with the keyword "#Horeg" for the period of January 1 to August 15, 2025, yielding 7,007 raw posts. Data labeling was performed using a semi-automated approach with the TextBlob lexicon, using polarity thresholds of $P > 0.05$ (positive), $P < -0.05$ (negative), and $-0.05 \leq P \leq 0.05$ (neutral). The labels were validated by an NLP expert (Hardian Oktavianto, S.Si., M.Kom.) on 700 stratified random samples (10% of the data), with a Cohen's Kappa agreement of $\kappa = 0.84$ (almost perfect agreement). The final label distribution consisted of 6,457 neutral reviews, 477 positive reviews, and 73 negative reviews. The text underwent five preprocessing stages (cleaning, tokenizing, stopword removal, normalization, and stemming) and was then represented using Term Frequency-Inverse Document Frequency (TF-IDF). To address the class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied until each class contained 6,457 samples. Evaluation was conducted using 5-Fold Cross-Validation on 70:30, 80:20, and 90:10 split ratios. The results show that the 80:20 ratio produced the best performance, with an accuracy of 99.98% and precision, recall, and F1-score values ranging from 0.99 to 1.00 across all sentiment classes. A qualitative analysis of TF-IDF feature weights confirmed that classification decisions were driven by keywords related to physical impact in the negative class (parah [severe], kesal [annoyed], marah [angry], bising [noisy]) and by words of appreciation in the positive class (mantap [great], senang [happy], suka [like], hiburan [entertainment]). This study provides a comprehensive and objective data-driven overview of public perception of sound Horeg.

Keywords: *Sentiment Analysis, Random Forest, TF-IDF, SMOTE, Sound Horeg*