INTISARI

Penelitian berjudul Penerapan Teknik Analisis Sentimen Isu Hak Cipta di Indonesia Menggunakan Algoritma random forest bertujuan untuk menganalisis dan mengklasifikasikan sentimen masyarakat Indonesia terkait isu hak cipta dengan membatasi variabel pada sentimen positif, negatif, dan netral. Data diperoleh dari 8.727 komentar di berbagai video Youtube melalui teknik Scraping. Proses pelabelan otomatis dilakukan terlebih dahulu menggunakan model DistilBERT multilingual yang menghasilkan distribusi sentimen negatif sebesar 55,77%, positif 42,57%, dan netral 1,66%. Setelah itu, data teks diproses melalui tahapan case folding, data cleaning, tokenisasi, stopword removal, dan stemming menggunakan Sastrawi stemmer. Untuk mengatasi masalah ketidakseimbangan data, digunakan teknik Synthetic Minority Oversampling Technique (SMOTE) sehingga performa model random forest tetap solid, dengan hasil prediksi benar 4.343 untuk kelas negatif dan 3.239 untuk kelas positif. Penelitian ini juga menghasilkan aplikasi web interaktif berbasis Streamlit yang mampu melakukan prediksi sentimen secara langsung dan menyediakan visualisasi data. Dengan demikian, penelitian ini memberikan kontribusi berupa pipeline analisis sentimen yang komprehensif, menggabungkan labeling berbasis transformer dengan ensemble learning, serta framework yang dapat digunakan kembali untuk isu serupa.

Kata kunci: analisis sentimen, hak cipta, random forest, komentar Youtube, Indonesia.

ABSTRACT

This research, entitled Implementation of Oversampling Techniques in Sentiment Analysis of Copyright Issues in Indonesia Using the Random Forest Algorithm aims to analyze and classify public sentiment in Indonesia regarding copyright issues, limited to positive, negative, and neutral sentiments. The dataset consists of 8,727 comments collected from various YouTube videos through a Scraping technique. Automatic labeling was first conducted using the DistilBERT multilingual model, which resulted in a sentiment distribution of 55.77% negative, 42.57% positive, and 1.66% neutral. Subsequently, the text data underwent preprocessing stages including case folding, data cleaning, tokenization, stopword removal, and stemming using the Sastrawi stemmer. To address the problem of class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied, ensuring that the Random Forest model maintained solid performance, with 4,343 correct predictions for the negative class and 3,239 for the positive class. This research also developed an interactive web application using Streamlit that enables real-time sentiment prediction and data visualization. Thus, this study contributes by providing a comprehensive sentiment analysis pipeline, combining transformer-based labeling with ensemble learning, and offering a reusable framework for analyzing public opinion on similar issues.

Keywords: sentiment analysis, copyright, Random Forest, youtube