ABSTRACT

Tangerang Selatan has been recorded as one of the cities with the worst air quality in Indonesia from 2023 to 2024. This condition has had a significant impact on public health. Public concern regarding this issue is often voiced through social media, particularly Twitter. This study aims to develop a sentiment analysis model based on machine learning and to apply non-parametric statistical approaches, including the chi-square test, Kruskal-Wallis test, and Eta test, to analyze the relationship between public sentiment and the Air Quality Index (AQI). The research process follows the KDD stages, including data selection, data preprocessing, data transformation, data mining, and evaluation. The Support Vector Machine (SVM) algorithm was used to classify sentiment into two categories, positive and negative, achieving an accuracy of 94.3% and an AUC value of 0.9497. The Net Sentiment Score (NSS) was then calculated to determine the daily sentiment tendency used in the relationship analysis. The chi-square test showed a χ^2 *statistic value of 27.757 with a p-value < 0.001, indicating a statistically significant relationship between sentiment and air quality. Meanwhile, the Kruskal-Wallis test produced an H value of 7.932 with a p-value of 0.047, indicating significant differences in sentiment scores across AQI categories. Additionally, the Eta test value of 0.059 suggests a moderate effect of air quality on sentiment. This study provides insights into public awareness of pollution impacts, public opinion as an early indicator of air quality conditions, and supports the development of Natural Language Processing in environmental issues.*

Keywords— ***Air Quality, Net Sentiment Score (NSS), Non-Parametric Test, Support Vector Machine (SVM), Sentiment Analysis.***