

ABSTRACT

Clean air is a fundamental human necessity. As Indonesia's economic center, Jakarta faces high levels of air pollution, posing serious health risks to its residents. Public opinion on this issue is widely expressed through both social media and mass media. Twitter (now X) enables users to voice spontaneous personal reactions, while online news portals present information in a more structured and editorialized format. This study adopts the Knowledge Discovery in Databases (KDD) framework to analyze public perception of Jakarta's air quality as conveyed on Twitter and in online news headlines. The dataset consists of 2,084 tweets and 535 news headlines collected throughout 2024 using keyword-based scraping. Sentiment analysis was conducted using the IndoBERT model, incorporating focal loss and data augmentation to handle class imbalance in the Twitter dataset, and class weighting for the news dataset. Evaluation results show accuracy scores of 87.8% for Twitter and 88.2% for news headlines. Net sentiment analysis revealed a strong negative tone on Twitter (-76.6%) and a more neutral tone in news media (+20.75%). Chi-Square testing showed no statistically significant association ($p = 0.3454$), with Cramér's $V = 0.14$ indicating a very weak correlation between the two platforms. To enrich the analysis, topic modeling using Latent Dirichlet Allocation (LDA) was applied. The results suggest that Twitter discourse tends to reflect emotional complaints and personal experiences, while news headlines emphasize institutional narratives and governmental responses.

Keywords: Air quality, Correlation, IndoBERT, Sentiment analysis, Topic modeling.