

ABSTRACT

A Comparative Analysis of BERTopic and G-BERTopic for Topic Modeling on Indonesian Short-Text News Headlines

By

IZAAZ RAHMAN AKBAR

21/472855/PA/20348

The rapid growth of short-text data such as news headlines poses significant challenges for topic modeling due to limited context and sparse word co-occurrence, particularly in the Indonesian news domain. This research compares BERTopic and G-BERTopic for topic modeling on Indonesian short-text news headlines from the COVID-19 period (2020–2022) using IndoSBERT embeddings. BERTopic integrates transformer embeddings with dimensionality reduction, clustering, and c-TF-IDF, while G-BERTopic uses pre-clustered knowledge to constrain topic formation. The models are evaluated using topic coherence and the Davies–Bouldin Index. Experimental results show that BERTopic outperforms G-BERTopic under practical topic limits, producing more coherent, better-separated topics with substantially lower computational cost. Sensitivity analysis further indicates that increasing the maximum topic limit in G-BERTopic narrows the performance gap, but substantially increases runtime. These findings suggest that BERTopic’s unconstrained clustering strategy is more suitable for large-scale short-text corpora with unknown topic structures.

Keywords – Topic Modeling, BERTopic, Indonesian News Headlines, Short-Text Analysis, Topic Coherence