

ABSTRAK

Pemilihan representasi teks menjadi salah satu faktor yang dapat memengaruhi kualitas topik yang dihasilkan oleh metode *Latent Dirichlet Allocation* (LDA). Permasalahan dalam penelitian ini adalah belum diketahuinya representasi teks yang menghasilkan kualitas topik terbaik antara *Bag of Words* (BoW) dan *Term Frequency–Inverse Document Frequency* (TF-IDF) pada komentar TikTok terkait Rancangan Undang-Undang (RUU) Perampasan Aset. Oleh karena itu, penelitian ini bertujuan untuk membandingkan nilai *topic coherence* yang dihasilkan oleh metode LDA menggunakan representasi BoW dan TF-IDF. Data yang digunakan berjumlah 2.954 komentar yang diperoleh melalui teknik *web scraping*. Tahapan penelitian meliputi *preprocessing teks*, pembentukan fitur *unigram* dan *bigram*, representasi teks menggunakan BoW dan TF-IDF, pemodelan topik menggunakan LDA, optimasi parameter menggunakan *Grid Search*, serta evaluasi model menggunakan *topic coherence*. *Grid Search* dilakukan dengan menguji variasi jumlah topik, alpha, dan beta. Model dilatih menggunakan *passes* sebesar 10, *iterations* sebesar 100, dan *random_state* sebesar 42. Hasil penelitian menunjukkan bahwa representasi BoW memperoleh *coherence score* tertinggi sebesar 0,5948 dengan jumlah topik 4, alpha 0,9, dan beta 0,8. Sementara itu, representasi TF-IDF memperoleh *coherence score* tertinggi sebesar 0,5334 dengan jumlah topik 6, alpha 1,0, dan beta 0,9. Berdasarkan hasil tersebut, representasi BoW menghasilkan nilai *topic coherence* lebih tinggi dibandingkan TF-IDF. Hasil interpretasi topik juga menunjukkan bahwa BoW menghasilkan pemisahan tema yang lebih jelas, yaitu topik mengenai RUU Perampasan Aset, aparat penegak hukum, DPR dan Puan Maharani, serta presiden, rakyat, dan aksi publik. Dengan demikian, representasi BoW menjadi representasi teks terbaik dalam penelitian ini berdasarkan nilai *topic coherence* dan interpretasi topik yang dihasilkan.

Kata Kunci: *Bag of Words*, *Latent Dirichlet Allocation*, TF-IDF, TikTok, *Topic Modeling*

ABSTRACT

Text representation selection is one of the factors that can influence the quality of topics generated by the Latent Dirichlet Allocation (LDA) method. The problem addressed in this study is that the text representation that produces the best topic quality between Bag of Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF) for TikTok comments related to the Asset Forfeiture Bill remains unknown. Therefore, this study aims to compare the topic coherence values generated by the LDA method using BoW and TF-IDF representations. The dataset used in this study consists of 2,954 comments obtained through web scraping. The research stages include text preprocessing, unigram and bigram feature formation, text representation using BoW and TF-IDF, topic modeling using LDA, parameter optimization using Grid Search, and model evaluation using topic coherence. Grid Search was conducted by testing variations in the number of topics, alpha, and beta. The model was trained using 10 passes, 100 iterations, and a random_state of 42. The results show that the BoW representation achieved the highest coherence score of 0.5948 with 4 topics, alpha of 0.9, and beta of 0.8. Meanwhile, the TF-IDF representation achieved the highest coherence score of 0.5334 with 6 topics, alpha of 1.0, and beta of 0.9. Based on these results, the BoW representation produced a higher topic coherence value than TF-IDF. The topic interpretation results also show that BoW produced clearer thematic separation, namely topics related to the Asset Forfeiture Bill, law enforcement officers, the House of Representatives and Puan Maharani, as well as the president, the public, and public actions. Thus, BoW is the best text representation in this study based on the topic coherence value and the resulting topic interpretation.

Keywords: *Bag of Words, Latent Dirichlet Allocation, TF-IDF, TikTok, Topic Modeling*