

ABSTRAK

Pertumbuhan pesat jumlah berita daring berbahasa Indonesia seiring kemajuan teknologi informasi telah menciptakan tantangan tersendiri dalam analisis manual, terutama karena volume data yang masif serta sifatnya yang tidak terstruktur. Oleh karena itu, diperlukan metode pemodelan topik untuk mengidentifikasi struktur tematik secara otomatis dari kumpulan dokumen teks. Penelitian ini bertujuan untuk membandingkan kinerja metode Latent Dirichlet Allocation (LDA) dan Non-Negative Matrix Factorization (NMF) dalam pemodelan topik pada berita online berbahasa Indonesia. Dataset yang digunakan terdiri atas 4.000 berita yang diperoleh melalui proses web scraping dari beberapa portal berita nasional dengan periode publikasi tahun 2020–2025, terbagi ke dalam empat kategori, yaitu politik, ekonomi, teknologi, dan olahraga. Tahapan penelitian meliputi preprocessing teks yang terdiri dari cleaning, case folding, tokenizing, stopword removal, dan stemming, dilanjutkan dengan representasi data menggunakan TF-IDF. Selanjutnya, dilakukan pencarian jumlah topik optimal dengan menguji variasi $K=2$ hingga $K=8$ berdasarkan nilai Topic Coherence. Hasil pengujian menunjukkan bahwa nilai C_v tertinggi untuk metode LDA diperoleh pada $K=5$ (0,6301), sedangkan untuk metode NMF diperoleh pada $K=3$ (0,7934). Dengan mempertimbangkan kesesuaian terhadap struktur empat kategori dataset (politik, ekonomi, teknologi, dan olahraga) serta untuk menghindari topik yang terlalu umum pada $K=3$ dan tumpang tindih pada $K=5$, penelitian ini menetapkan $K=4$ sebagai jumlah topik optimal pada pemodelan akhir, yang tetap menghasilkan nilai Coherence yang tinggi (LDA=0,6012; NMF=0,7645). Hasil perbandingan pada $K=4$ menunjukkan bahwa NMF menghasilkan rata-rata Topic Coherence sebesar 0,7645, jauh lebih tinggi dibandingkan LDA yang hanya mencapai 0,4830 (peningkatan sebesar 0,2815). Dengan demikian, berdasarkan metrik Topic Coherence, NMF terbukti lebih efektif dibandingkan LDA dalam pemodelan topik pada berita online berbahasa Indonesia.

Kata Kunci: Berita online, LDA, NMF, Pemodelan Topik, Topic Coherence

ABSTRACT

The rapid growth of Indonesian-language online news along with advances in information technology has created challenges in manual analysis, particularly due to the massive volume and unstructured nature of the data. Therefore, topic modeling methods are needed to automatically identify thematic structures from text document collections. This study aims to compare the performance of the Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) methods in topic modeling of Indonesian-language online news. The dataset consists of 4,000 news articles obtained through web scraping from several national news portals with publication periods from 2020 to 2025, divided into four categories: politics, economics, technology, and sports. The research stages include text preprocessing consisting of cleaning, case folding, tokenizing, stopword removal, and stemming, followed by data representation using TF-IDF. Furthermore, the optimal number of topics was determined by testing variations of $K=2$ to $K=8$ based on Topic Coherence scores. The results showed that the highest C_v for LDA was obtained at $K=5$ (0.6301), while for NMF it was obtained at $K=3$ (0.7934). Considering the alignment with the four-category dataset structure (politics, economics, technology, and sports) and to avoid topics that are too general at $K=3$ or overlapping at $K=5$, this study established $K=4$ as the optimal number of topics for the final modeling, which still yielded high C_v values (LDA=0.6012; NMF=0.7645). The comparison at $K=4$ shows that NMF achieved an average Topic Coherence of 0.7645, significantly higher than LDA which only reached 0.4830 (an improvement of 0.2815). Thus, based on the Topic Coherence metric, NMF is proven to be more effective than LDA for topic modeling of Indonesian-language online news.

Keywords: Online news, LDA, NMF, Topic Modeling, Topic Coherence.