INTISARI

Pemodelan Topik Menggunakan *Latent Dirichlet Allocation* dan *Sentence-Bidirectional Encoder Representation from Transformer* Pada Laman Berita *Hoax*

Oleh

Vitadevi Rahmawati

20/455525/PA/19740

Pada saat ini, banyak informasi palsu yang dikemas dalam bentuk pemberitaan melalui berbagai media massa mengikuti *trend* pembicaraan pada saat itu sehingga akan lebih banyak menarik masyarakat untuk mempercayainya. Pemodelan topik merupakan salah satu teknik analisis data teks yang bermanfaat dalam menemukan makna tersebunyi dalam dokumen. Penentuan topik – topik yang ada pada laman berita *hoax* dapat membantu masyarakat untuk lebih berhati – hati ketika mendapat suatu informasi. *Latent Dirichlet Allocation* (LDA) merupakan model probabilitas generatif yang digunakan dalam pemodelan topik dan merupakan salah satu metode klasik yang hingga saat ini masih banyak digunakan. Di sisi lain, *Sentence-Bidirectional Encoder Representation from Transformer* (SBERT) adalah model pembelajaran mesin *modern* yang didasarkan pada arsitektur *transformer* dan mempertimbangkan kata – kata di sekitarnya dari dua sisi. Penelitian ini, menggunakan studi kasus berupa berita *hoax* mulai 1 Januari sampai 31 Agustus 2023 yang didapat dari laman berita *hoax* turnbackhoax.id. Teknik pemodelan topik yang digunakan adalah LDA dan SBERT yang mana pada LDA digunakan dua teknik representasi teks, yaitu *Bag of Words* (BoW) dan *Term Frequency-Inverse Document Frequency* (TF-IDF). Diperoleh model terbaik berdasarkan hasil evaluasi model, yaitu SBERT dengan nilai koherensi sebesar 0,649. Topik yang didapatkan adalah sebanyak 8 topik, yaitu: Topik 1: Konten Video Palsu Mengenai Calon Presiden Indonesia (919 berita), Topik 2: Akun Sosial Media Palsu Yang Mengatasnamakan Pemerintah (97 berita), Topik 3: Manipulasi Fakta Dan Foto Palsu Yang Berkaitan Dengan Akun Tertentu (86 berita), Topik 4: Piala Dunia Speak Bola (44 berita), Topik 5: Facebook Banyak Memberikan Artikel *Hoax* Mengenai Pemerintahan Indonesia (37 berita), Topik 6: Penyakit Dan Penyebabnya (31 berita), Topik 7: Berita *Hoax* Banyak Berisi Video Dan Gambar Berupa Potongan – Potongan (29 berita), dan Topik 8: Vaksin Untuk Covid-19 (26 berita).

Kata kunci: pemodelan topik, *Latent Dirichlet Allocation* (LDA), *Bidirectional Encoder Representation from Transformer* (BERT), berita *hoax, sentence embedding*

ABSTRACT

Topic modeling Using Latent Dirichlet Allocation and *Sentence-Bidirectional Encoder Representations from Transformers on Hoax News Page*

By

Vitadevi Rahmawati

20/455525/PA/19740

A significant amount of false information is packaged in the form of news through various mass media, following the prevailing trends of conversation, to attract more people into believing it. Topic modelling serves as one of the valuable text data analysis techniques to uncover hidden meanings within documents. Identifying the topics within hoax news articles can assist the public in being more cautious when receiving information. Latent Dirichlet Allocation (LDA) stands as a generative probability model used in topic modelling and remains one of the classical methods that is widely employed even today. In contrast, Sentence-Bidirectional Encoder Representation from Transformer (SBERT) is a modern machine learning model based on the transformer architecture, considering words from both sides. In this study, a case study was conducted involving hoax news articles from January 1 to August 31, 2023, obtained from the hoax news website turnbackhoax.id. The topic modeling techniques used are LDA and SBERT, with LDA employing two text representation techniques, Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). The best model was determined based on model evaluation, with SBERT achieving a coherence value of 0.649. A total of eight topics were identified, which are as follows: Topic 1: Fake Video Content About the Indonesian Presidential Candidate (919 articles), Topic 2: Fake Social Media Accounts Impersonating the Government (97 articles), Topic 3: Manipulation of Facts and Fake Photos Related to Specific Accounts (86 articles), Topic 4: Speak Bola World Cup (44 articles), Topic 5: Facebook Contains Many Hoax Articles About the Indonesian Government (37 articles), Topic 6: Diseases and Their Causes (31 articles), Topic 7: Hoax News Often Contain Videos and Image Fragments (29 articles), and Topic 8: COVID-19 Vaccines (26 articles).

Keywords: *topic modelling, Latent Dirichlet Allocation* (LDA), *Bidirectional Encoder Representation from Transformer* (BERT), *hoax news, sentence embedding*