

IMPLEMENTASI NORMALISASI TEKS INTEGRATIF BERBASIS LEKSIKON DAN WORD EMBEDDING *FASTTEXT* PADA PEMODELAN TOPIK ISU INDONESIA GELAP DI PLATFORM X

Oleh

Fransesko Indrajid, NIM 2115091048

Jurusan Teknik Informatika

ABSTRAK

Fenomena "IndonesiaGelap" mencuat pada awal tahun 2025 sebagai respons publik terhadap Instruksi Presiden Nomor 1 Tahun 2025 dan menjadi simbol keresahan kolektif masyarakat terhadap kondisi sosial-politik nasional di platform X. Tingginya intensitas percakapan yang dinamis dan penggunaan bahasa tidak baku (*slang*) menjadi tantangan utama dalam mengekstraksi informasi dari data media sosial tersebut. Penelitian ini bertujuan untuk memodelkan topik isu "IndonesiaGelap" menggunakan algoritma *Latent Dirichlet Allocation* (LDA) dengan menerapkan strategi normalisasi teks integratif. Pendekatan ini menggabungkan leksikon gabungan dan *word embedding* FastText untuk menangani permasalahan *sparsity* dan kata-kata *out-of-vocabulary* (OOV). Dari total 57.533 data mentah yang dikumpulkan, sebanyak 48.735 *tweet* digunakan dalam pemodelan setelah melalui tahap prapemrosesan. Hasil penelitian berhasil mengidentifikasi enam topik dominan, yaitu: (1) Kebijakan, Program, dan Revisi UU, (2) Sumber Daya, Korupsi, Pajak, dan Penderitaan Rakyat, (3) Pemilu, Pemimpin, Kekuasaan, dan Demokrasi, (4) Aksi Mahasiswa dan Wacana Publik, (5) Dukungan Moral dan Solidaritas Massa, serta (6) Kecaman Moral dan Religius terhadap Aparat. Evaluasi model menunjukkan bahwa skenario normalisasi integratif menghasilkan kinerja terbaik dengan skor *Topic Coherence* tertinggi sebesar 0.4770 dan *Perplexity* yang stabil pada 1388.25. Validitas hasil diperkuat melalui evaluasi berbasis manusia (*human-centered evaluation*), yang mencatatkan akurasi *Word Intrusion* sebesar 92.22% dan *Topic Intrusion* sebesar 86.67%. Capaian ini membuktikan bahwa integrasi leksikon dan FastText efektif dalam meningkatkan kualitas representasi topik pada data informal, menghasilkan wacana yang koheren, distinktif, dan dapat diinterpretasikan secara logis sesuai konteks isu yang sebenarnya.

Kata Kunci: IndonesiaGelap, *Latent Dirichlet Allocation*, *Topic Modeling*, Normalisasi Teks, *FastText*, Leksikon.

IMPLEMENTATION OF INTEGRATIVE TEXT NORMALIZATION USING LEXICON AND FASTTEXT WORD EMBEDDING IN TOPIC MODELING OF THE 'INDONESIA GELAP' ISSUE ON PLATFORM X

By

Fransesko Indrajid, NIM 2115091048

Department of Informatics Engineering

ABSTRACT

The "IndonesiaGelap" phenomenon emerged in early 2025 as a public response to Presidential Instruction Number 1 of 2025, symbolizing collective societal unrest regarding national socio-political conditions on the X platform. The high intensity of dynamic conversations and the prevalence of non-standard language (slang) pose significant challenges in extracting information from such social media data. This study aims to model topics surrounding the "IndonesiaGelap" issue using the Latent Dirichlet Allocation (LDA) algorithm by implementing an integrative text normalization strategy. This approach combines a Merged Lexicon and FastText word embedding to address sparsity and out-of-vocabulary (OOV) issues. From a total of 57,533 raw data collected, 48,735 tweets were utilized for modeling after preprocessing. The study successfully identified six dominant topics: (1) Policies, Programs, and Law Revisions; (2) Resources, Corruption, Taxes, and People's Suffering; (3) Elections, Leaders, Power, and Democracy; (4) Student Actions and Public Discourse; (5) Moral Support and Mass Solidarity; and (6) Moral and Religious Condemnation of Apparatus. Model evaluation demonstrated that the integrative normalization scenario yielded the best performance, achieving the highest Topic Coherence score of 0.4770 and a stable Perplexity of 1388.25. The validity of the results was further reinforced through human-centered evaluation, recording a Word Intrusion accuracy of 92.22% and Topic Intrusion accuracy of 86.67%. These findings prove that the integration of lexicon and FastText is effective in enhancing topic representation quality in informal data, resulting in coherent, distinctive, and logically interpretable discourse aligned with the actual context of the issue.

Keywords: IndonesiaGelap, Latent Dirichlet Allocation, Topic Modeling, Text Normalization, FastText, Lexicon.