BAB I

PENDAHULUAN

1.1 Latar Belakang

Pada awal tahun 2025, “IndonesiaGelap” mencuat sebagai isu yang ramai diperbincangkan di platform X, dimulai sejak 17 Februari 2025 ketika respons publik dimulai terhadap terbitnya Instruksi Presiden Nomor 1 Tahun 2025 yang memangkas anggaran pendidikan demi mendanai program Makan Bergizi Gratis (MBG) (Clairine et al., 2025). Ramai dibicarakan dalam kurun waktu hampir dua bulan, “IndonesiaGelap” tidak hanya menjadi *trending*, tetapi juga berubah menjadi simbol dari keresahan kolektif masyarakat terhadap kondisi sosial-politik nasional, mencakup kritik terhadap militerisasi sipil, ketimpangan ekonomi, kemunduran demokrasi, serta ketidakpuasan terhadap kinerja pemerintahan yang baru menjabat (A. M. Nur et al., 2025). Mengingat tingginya intensitas opini dan percakapan yang muncul, diperlukan pendekatan yang komprehensif dan berbasis data untuk mengidentifikasi topik-topik utama dalam wacana ini, guna memperoleh pemahaman yang lebih mendalam mengenai dinamika isu dan alasan di balik meluasnya perhatian publik terhadap isu ini.

Salah satu pendekatan yang bisa digunakan untuk memahami isu #IndonesiaGelap secara lebih mendalam adalah dengan memanfaatkan salah satu cabang kecerdasan buatan yang dikenal sebagai Natural Language Processing (NLP). NLP atau pemrosesan bahasa alami secara umum berfungsi untuk memungkinkan komputer memahami, memproses, dan menghasilkan bahasa manusia secara alami (Egger & Gokce, 2022). Dalam hal ini, penerapan NLP yang tepat digunakan adalah pemodelan topik, yakni metode yang mampu mengekstraksi informasi tematik dari kumpulan teks dalam jumlah besar tanpa perlu pelabelan manual *(unsupervised)* (Yin & Yuan, 2022). Menggunakan model dengan pendekatan probabilistik seperti *Latent Dirichlet Allocation* (LDA), tema atau topik dalam teks mampu diidentifikasi secara efisien bahkan pada data yang tidak terstruktur (Ozyurt et al., 2024). LDA menghasilkan topik-topik yang mudah

diinterpretasikan, sehingga cocok untuk menganalisis percakapan publik di media sosial yang bersifat dinamis, masif, dan penuh variasi bahasa (Ariansyah & Indahyanti, 2024; Nurhaliza, 2024). Oleh karena itu, model ini relevan untuk digunakan dalam memahami substansi dan topik-topik yang berkembang di balik tagar #IndonesiaGelap, yang mencerminkan opini dan respons masyarakat.

Namun, penelitian NLP yang menggunakan media sosial sebagai sumber data utama kerap menghadapi tantangan berupa teks yang tidak baku, penuh singkatan, slang, serta kesalahan penulisan, yang menyulitkan model dalam memahami makna secara akurat (Raif et al., 2024). Temuan Purwitasari et al., (2023) menunjukkan bahwa lima kata teratas yang tidak bisa diklasifikasi dengan benar adalah kata-kata informal. Sementara itu, Halimi et al., (2021) juga menemukan permasalahan serupa dan merekomendasikan perlunya pendekatan yang lebih efektif untuk menangani konversi kata informal ke bentuk formal. Dari beberapa media sosial yang tersedia, platform X dipilih sebagai objek penelitian karena sebagian besar penelitian NLP yang masih mengalami kendala dalam pemrosesan bahasa informal menggunakan data yang bersumber dari X, ditemukan bahwa hal ini bisa terjadi karena platform X hanya mengizinkan 280 karakter per unggahan, sehingga pengguna terdorong menyingkat kata dan menggunakan ekspresi informal ketika mengunggah *tweet* (Lee & Song, 2022; Meddeb & Romdhane, 2022). Selain itu, platform ini juga bersifat terbuka, cepat, dan aktif digunakan untuk mengekspresikan opini publik secara real-time, sehingga memungkinkan penelusuran persepsi masyarakat terhadap isu-isu aktual seperti #IndonesiaGelap. Lebih lanjut, X telah terbukti efektif digunakan dalam berbagai studi analisis sentimen dan opini publik (Pramayasa et al., 2023), menjadikannya platform yang sesuai untuk penelitian berbasis NLP terhadap teks informal.

Untuk mengatasi permasalahan akibat teks informal, maka dilakukan tahapan normalisasi sebelum pemrosesan lebih lanjut. Salah satu metode yang umum digunakan adalah normalisasi berbasis leksikon, yaitu dengan memanfaatkan daftar kata tidak baku beserta padanan formalnya. Penerapan leksikon seperti *Colloquial Indonesian Lexicon* dalam proses normalisasi data Twitter untuk keperluan analisis sentimen terbukti menghasilkan peningkatan kualitas representasi serta akurasi yang didapatkan oleh model (Aisy & Prasetiyo, 2023; Sanjaya et al., 2022).

Leksikon serupa yaitu *IndoCollex* juga diterapkan oleh Bustamin, (2025), dan berhasil meningkatkan akurasi klasifikasi sentimen sebesar 3,47% . Meskipun cukup efektif, pendekatan leksikon ternyata memiliki keterbatasan yaitu cakupan kosakata. Kata-kata baru, variasi ejaan, atau istilah yang tidak tercantum dalam daftar leksikon sering kali luput dari proses normalisasi. Untuk menutupi kelemahan ini, pendekatan berbasis *word embedding* dapat digunakan sebagai pelengkap. Dengan merepresentasikan kata dalam bentuk vektor pada ruang embedding, sistem NLP dapat mengenali kemiripan semantik antar kata, bahkan ketika terdapat perbedaan ejaan atau bentuk morfologis (R. Nur et al., 2025). Salah satu model yang banyak digunakan dalam konteks ini adalah *FastText*, Ardinata et al., (2024) dan R. Nur et al., (2025) menunjukkan bahwa *FastText* efektif dalam meningkatkan akurasi klasifikasi teks pada data media sosial karena kemampuannya mengenali variasi bentuk kata dan ejaan tidak baku, termasuk kata-kata baru atau tidak umum, melalui representasi sub-kata dalam ruang vektor. Selain lebih adaptif terhadap data informal, *FastText* juga menunjukkan efisiensi waktu dalam proses normalisasi, sehingga sangat sesuai untuk diterapkan pada teks pendek dan dinamis seperti tweet.

Berdasarkan uraian yang telah dipaparkan, maka penelitian ini dirancang dengan tujuan untuk menggali pengetahuan terkait wacana digital isu *"IndonesiaGelap"*. Untuk mendukung tujuan tersebut, penelitian ini menggunakan pendekatan pemodelan topik dalam kerangka NLP, dengan terlebih dahulu mengoptimalkan representasi teks melalui proses normalisasi. Pendekatan normalisasi yang digunakan bersifat integratif, yakni dengan menggabungkan metode berbasis leksikon dan word embedding. Leksikon yang akan digunakan ialah *Colloquial Indonesian Lexicon* dan *IndoCollex,* sementara model word embedding *FastText* dimanfaatkan untuk mengenali dan menyesuaikan kata-kata informal di luar leksikon. Dengan pendekatan ini, penelitian yang berjudul ***"Implementasi Normalisasi Teks Integratif Berbasis Leksikon dan Word Embedding FastText pada Pemodelan Topik Isu IndonesiaGelap di Platform X"*** diharapkan mampu mengungkap struktur topik yang tersembunyi dalam diskursus digital secara lebih akurat serta memberikan pemahaman yang lebih mendalam terhadap dinamika isu yang berkembang di masyarakat.

1.2 Rumusan Masalah

Dari pembahasan latar belakang yang telah disampaikan, dapat disimpulkan permasalahan yang menjadi titik perhatian utama dalam penelitian ini, yaitu:

1. Perlunya pendekatan yang lebih adaptif dalam menangani teks informal, seperti penggunaan *slang*, singkatan, dan kesalahan penulisan, yang umum ditemukan dalam data media sosial.
2. Perlunya metode yang tepat untuk mengidentifikasi topik-topik yang membentuk wacana *#IndonesiaGelap* di platform X.

Berdasarkan permasalahan penelitian tersebut, maka pertanyaan penelitian dalam penelitian ini adalah sebagai berikut:

1. Bagaimana implementasi normalisasi teks integratif berbasis leksikon dan *word embedding FastText* dalam melakukan normalisasi teks pada proses pemodelan topik isu *IndonesiaGelap* di platform X?
2. Bagaimana hasil dan evaluasi model LDA dalam pemodelan topik isu IndonesiaGelap pada platform X?

1.3 Tujuan Penelitian

Berikut adalah tujuan penelitian berdasarkan rumusan masalah yang telah dirumuskan sebelumnya:

1. Untuk mengimplementasikan metode normalisasi teks integratif berbasis leksikon dan *word embedding FastText* dalam proses pemodelan topik isu *IndonesiaGelap* di platform X.
2. Untuk mengetahui hasil dan mengevaluasi model LDA dalam pemodelan topik isu *IndonesiaGelap* di platform X.

1.4 Ruang Lingkup Penelitian

Guna membuat penelitian yang akan dilakukan menjadi lebih terarah, maka berikut merupakan ruang lingkup penelitian dalam penelitian ini:

1. Dataset yang digunakan dalam penelitian ini adalah tweet dari platform X (sebelumnya Twitter) yang menggunakan kata kunci “indonesiagelap”. Data dikumpulkan dalam rentang waktu Februari hingga Agustus 2025, berfokus

pada periode ketika #IndonesiaGelap sedang *trending,* sehingga dapat merepresentasikan dinamika percakapan publik pada pembahasan isu tersebut.

2. Dataset yang digunakan merupakan *tweet* yang berbahasa Indonesia dan campuran antara bahasa Indonesia dan bahasa lainnya, tetapi bahasa lainnya akan dihapus nantinya pada tahap praproses agar fokus penelitian lebih terarah.
3. Kata informal dengan penulisan yang sama atau mirip pada leksikon, namun memiliki lebih dari satu padanan kata formal, akan ditangani melalui tahapan secara manual pada tahapan normalisasi. Pemilihan padanan kata dilakukan secara manual dengan mempertimbangkan konteks tweet., seperti kata *"km"* dapat merujuk pada *"kilometer"* sebagai singkatan, atau *"kami"* maupun *"kamu"* sebagai bentuk slang.

1.5 Manfaat Penelitian

Berdasarkan rumusan masalah dan tujuan penelitian yang telah diuraikan sebelumnya, berikut manfaat yang diharapkan dapat diperoleh melalui penelitian ini:

1. Bagi Pemerintah dan Pejabat Publik

a. Memberikan gambaran mengenai perkembangan isu sosial di media sosial, khususnya yang terkait dengan tagar *#IndonesiaGelap*, yang dapat dijadikan rujukan dalam memahami opini publik.

b. Dapat digunakan sebagai bahan pertimbangan dalam merumuskan kebijakan publik atau strategi komunikasi pemerintah yang lebih tepat sasaran dan berbasis data.

2. Bagi Pihak yang Ingin Menyebarluaskan Informasi Isu IndonesiaGelap

a. Menyediakan hasil pemodelan topik yang menampilkan daftar topik utama beserta kata kunci relevan, sehingga pihak terkait dapat mengenali isu yang sedang banyak diperbincangkan di media sosial secara cepat dan terukur.

b. Memberikan peta persebaran dan arah percakapan publik yang telah diinterpretasikan secara tematik, sehingga dapat dijadikan dasar untuk merancang narasi dan menyampaikan informasi yang relevan, utuh, dan tepat sasaran sesuai minat dan perhatian publik.

3. Bagi Pembaca

a. Memberikan informasi mengenai topik-topik yang berkembang di media sosial terkait isu *#IndonesiaGelap*, sehingga pembaca dapat memahami isu secara lebih menyeluruh.

b. Menyediakan wawasan mengenai pentingnya proses normalisasi teks dalam analisis media sosial berbahasa Indonesia, khususnya yang mengandung bahasa informal.

c. Dapat dijadikan referensi bagi pembaca atau peneliti yang ingin melakukan kajian serupa di masa mendatang, baik dalam konteks pemodelan topik maupun analisis media sosial.

4. Bagi Peneliti

a. Menambah wawasan dan pengalaman dalam mengembangkan metode pemodelan topik serta mengatasi tantangan normalisasi teks informal dalam penelitian NLP.

b. Menjadi sarana untuk mengimplementasikan ilmu dan keterampilan yang telah diperoleh selama menempuh studi pada Program Studi Sistem Informasi di Universitas Pendidikan Ganesha.