BAB I
PENDAHULUAN

1.1. Latar Belakang

Perkembangan jumlah pengguna media sosial menunjukkan tren yang signifikan dalam beberapa tahun terakhir. Terdapat hampir 4,95 miliar pengguna media sosial, dengan lebih dari separuh populasi dunia terlibat dalam media sosial (Alsemaree dkk., 2024). Berdasarkan laporan *We Are Social* tahun 2024, lebih dari 200 juta pengguna aktif media sosial berada di Indonesia. Angka tersebut mencerminkan pertumbuhan yang pesat, karena berdasarkan data pada tahun 2021, jumlah pengguna media sosial tercatat sekitar 170 juta (Tannia dan Monika, 2022). Peningkatan jumlah pengguna media sosial ini memberi pengaruh pada dunia usaha, dimana besar peluang bagi *marketing* dan *brand manager* untuk bekerja sama dengan konsumen guna meningkatkan visibilitas merek (Schivinski dan Dabrowski, 2016). Media sosial memiliki peran strategis sebagai wadah utama untuk menyalurkan opini, keluhan, dan pengalaman pelanggan di mana saja dan kapan saja (Jha, 2019). Media sosial juga mendorong keterlibatan pelanggan melalui kepuasan, emosi positif, dan kepercayaan (Hajek dkk., 2021).

Platform media sosial yang sering digunakan dalam mengekspresikan opini pelanggan adalah X (Twitter). Platform ini memiliki karakteristik unik dengan format teks pendek yang memungkinkan pengguna berbagi pengalaman dengan cepat dan mudah (Alipour dkk., 2023). X (Twitter) memproduksi sekitar 500 juta tweet yang diposting setiap bulan dari total 313 juta pengguna aktif (Yadav dkk., 2023). Kekayaan data media sosial tersebut telah membuka jalan baru bagi penelitian untuk mendapatkan wawasan tentang perilaku dan pengalaman manusia (Egger dan Yu, 2022). Dengan volume data opini pelanggan yang besar, media sosial X dapat menjadi sumber informasi yang berharga bagi perusahaan dalam memahami kepuasan pelanggan, mengidentifikasi masalah layanan, serta merancang strategi perbaikan layanan berbasis data. Namun, untuk dapat memanfaatkan data ini secara optimal, dibutuhkan pendekatan analitis yang efektif dalam mengekstraksi informasi dari teks yang singkat dan tidak terstruktur.

Pendekatan yang dapat digunakan untuk menganalisis volume besar opini pelanggan di media sosial adalah pemodelan topik. Metode pemodelan topik efektif untuk menyimpulkan topik-topik laten dari teks-teks pendek, yang secara khusus relevan untuk analisis opini di media sosial (Alsmadi dkk., 2024). Teknik ini memungkinkan pengelompokan otomatis dari opini yang tersebar menjadi topik-topik utama yang dapat diinterpretasikan (Wang dkk., 2023). Teknik pemodelan topik memfasilitasi ekstraksi wawasan yang bermakna dari teks dalam jumlah besar, memungkinkan pemahaman yang lebih dalam tentang perilaku dan pengalaman manusia yang tercermin dalam data *tweet* di Twitter (Egger dan Yu, 2022). Metode seperti *Latent Dirichlet Allocation* (LDA) telah banyak digunakan untuk mengidentifikasi struktur topik dalam teks besar yang merupakan salah satu metode penting dalam penambangan teks. LDA memungkinkan penemuan informasi laten yaitu informasi yang tidak bisa langsung dilihat atau diakses dengan mudah, tetapi memerlukan metode atau alat khusus untuk mengungkapnya, serta mengungkap hubungan antara dokumen yang tidak berurutan (Bastani dkk., 2019; Jelodar dkk., 2019). Namun sifat alami kalimat pendek seperti pada data *tweet* adalah memiliki struktur yang berantakan, menerapkan model topik konvensional seperti LDA dan pLSA secara langsung terbukti tidak efisien (Hajjem dan Latiri, 2017). Algoritma *Gibbs Sampling Dirichlet Mixture Model* (GSDMM) dibuat secara khusus untuk mengatasi keterbatasan dalam pemodelan topik teks pendek dengan cara mengelompokkan dokumen ke dalam sejumlah topik kecil yang lebih koheren tanpa memerlukan jumlah topik yang telah ditentukan sebelumnya (Yin dan Wang, 2014). Penelitian sebelumnya menunjukkan bahwa GSDMM memiliki keunggulan dibandingkan LDA dalam menangani dataset teks pendek dengan sparse data dan tingkat noise yang tinggi (Qiang dkk., 2022a). Dalam konteks media sosial, GSDMM lebih efektif dalam mengelompokkan tweet berdasarkan kemiripan topik, sehingga dapat memberikan hasil yang lebih akurat dalam analisis opini pelanggan (Weisser dkk., 2023).

Dalam analisis topik berbasis teks pendek seperti tweet, representasi data memainkan peran krusial dalam menentukan kualitas yang dihasilkan oleh model. Pendekatan awal terhadap representasi teks dalam analisis teks pendek sangat

bergantung pada model *bag-of-words* (BoW) (De Santis dkk., 2024; Hajek dkk., 2021). Model-model ini, meskipun sederhana untuk diimplementasikan, memiliki beberapa keterbatasan. Model-model tersebut mengabaikan urutan kata dan informasi kontekstual, yang menyebabkan hilangnya makna semantik dan berkurangnya akurasi, terutama pada teks-teks pendek di mana konteks sangat penting. Perkembangan teknik *embedding* kata, seperti Word2Vec dan GloVe, menandai pergeseran paradigma yang signifikan (De Santis dkk., 2024; Romero dkk., 2024; Si dkk., 2019). Teknik-teknik ini merepresentasikan kata-kata sebagai vektor dalam ruang dimensi tinggi, menangkap hubungan semantik antara kata-kata berdasarkan kemunculannya dalam dokumen / korpus besar. Hal ini memungkinkan penggabungan informasi kontekstual, yang mengarah pada peningkatan kinerja dalam tugas-tugas seperti klasifikasi teks dan analisis sentimen. Sebagai contoh, terdapat penelitian yang membandingkan model BoW dengan teknik *embedding* modern, dimana teknik *embedding* modern seperti BERT memiliki kinerja yang lebih unggul dalam mengklasifikasikan pengguna dalam diskusi media sosial medis (De Santis dkk., 2024).

Pengembangan *embedding* kata yang dikontekstualisasikan, seperti yang dihasilkan oleh BERT, ELMo, dan model berbasis transformator lainnya, semakin meningkatkan kemampuan teknik *embedding* (De Santis dkk., 2024; Si dkk., 2019). Model-model ini menangkap makna kata yang bergantung pada konteks, mengatasi keterbatasan penyematan statis. Sebagai contoh, kata “bank” dapat memiliki arti yang berbeda tergantung pada kata-kata di sekitarnya, dan *embedding* yang dikontekstualisasikan dapat menangkap nuansa ini. Penggunaan teknik BERT telah dieksplorasi dalam menganalisis proses pengambilan keputusan mahasiswa teknik, yang menunjukkan keefektifannya dalam menangkap topik-topik yang bernuansa dalam teks argumentatif (Romero dkk., 2024). Selain itu, penelitian telah mengeksplorasi penerapan *embedding* kata dengan teknik lain untuk meningkatkan kinerja. Menggabungkan *embedding* dengan sentimen berbasis leksikon dan indikator emosi serta pemodelan topik menunjukkan efek sinergis dan memiliki hasil yang baik (Hajek dkk., 2021). Terdapat juga penelitian yang menggunakan *embedding* FastText dan BERT untuk meningkatkan hasil pengelompokan otomatis

proposal proyek, menunjukkan penerapan teknik-teknik ini dalam berbagai domain (Aksoy dkk., 2023). Dalam konteks bahasa Indonesia, telah dikembangkan model IndoBERTweet, yang dilatih secara khusus menggunakan 409 juta token dari tweet berbahasa Indonesia. Hal tersebut membuatnya mampu menangkap nuansa bahasa yang digunakan dalam media sosial, termasuk variasi bahasa informal, singkatan, dan istilah spesifik yang sering digunakan oleh pengguna X di Indonesia IndoBERTweet bekerja dengan menghasilkan representasi vektor berdimensi tinggi untuk setiap kata atau dokumen dengan mempertimbangkan konteks penggunaannya, sehingga mampu mengenali sinonim, frasa idiomatik, serta nuansa kata dalam struktur informal seperti tweet. (Koto dkk., 2021).

Penelitian tentang penerapan *Gibbs Sampling Dirichlet Mixture Model* (GSDMM) yang dioptimasi dengan evaluasi semantik IndoBERTweet untuk analisis topik teks pendek opini pelanggan di media sosial X perlu dilakukan karena metode pemodelan topik konvensional seperti Latent Dirichlet Allocation (LDA) sering mengalami keterbatasan dalam menangani teks pendek dan noise seperti tweet (Hajjem dan Latiri, 2017). GSDMM merupakan metode yang lebih cocok untuk analisis teks pendek karena mampu mengelompokkan data dengan distribusi kata yang lebih terbatas dengan mengasumsikan bahwa setiap tweet hanya berkaitan dengan satu topik utama (Weisser dkk, 2023; Yin dan Wang, 2014). Namun untuk meningkatkan representasi semantik yang lebih dalam sesuai karakter bahasa informal yang sering digunakan di media sosial, Word2Vec atau GloVe, yang umumnya digunakan untuk representasi kata, tidak dapat menangani dengan baik kata-kata yang terkait dengan konteks yang digunakan dalam percakapan sehari-hari di media sosial. Penggunaan embedding teks yang dipra-latih bersama dengan pemodelan topik dapat menghasilkan interpretasi yang lebih baik dari topik yang muncul dalam teks (De Santis dkk., 2024). Oleh karena itu, penelitian ini mengusulkan pendekatan optimisasi dua fase. Fase pertama menggunakan GSDMM untuk membentuk berbagai kandidat model topik melalui eksplorasi kombinasi hyperparameter, fase kedua memanfaatkan IndoBERTweet sebagai evaluator eksternal untuk mengukur kualitas semantik topik-topik tersebut.

Penelitian ini sekaligus mengisi celah empiris mengenai efektifitas optimasi GSDMM dengan evaluasi semantik IndoBERTweet dalam menganalisis teks pendek opini pelanggan yang diperoleh dari media sosial X (Twitter). Data ini terdiri dari tweet yang membahas *provider* telekomunikasi di Indonesia, yang dikumpulkan menggunakan teknik *web scraping* dengan bantuan fungsi *Advanced Search* yang disediakan media sosial X untuk mengekstraksi hasil pencarian tweet secara otomatis berdasarkan kata kunci, rentang waktu, bahasa, dan parameter lain yang dapat ditentukan. Dari segi manfaat praktis, hasil penelitian ini dapat digunakan untuk mengembangkan alat analisis topik yang lebih akurat bagi bisnis di Indonesia, membantu mereka memahami opini pelanggan dengan lebih baik. Selain itu, penelitian ini juga dapat berkontribusi pada pengembangan *Natural Language Processing* (NLP) untuk bahasa Indonesia, dengan memperkenalkan metode alternatif dalam ekstraksi informasi dari media sosial.

1.2. Tujuan Penelitian

Penelitian ini bertujuan untuk membangun model analisis topik yang mengatasi keterbatasan model topik konvensional dalam mengolah teks pendek di media sosial. Untuk mencapai tujuan tersebut, penelitian ini menerapkan algoritma *Gibbs Sampling Dirichlet Mixture Model* (GSDMM) dengan variasi kombinasi *hyperparameter* untuk menangani karakteristik *sparsity* pada data teks pendek dan mengoptimalkan pemilihan model terbaik. Pendekatan ini digunakan untuk menyeleksi model yang memiliki kualitas topik paling optimal (koheren dalam satu topik dan terpisah jelas antar topik) dari kandidat model yang dihasilkan GSDMM. Model terbaik yang dihasilkan diharapkan mampu meningkatkan kualitas semantik dan menghasilkan topik dengan interpretabilitas yang tinggi.

1.3. Manfaat Penelitian

Penelitian ini diharapkan memberikan kontribusi baik secara teoritis maupun praktis. Secara teoritis, penelitian ini akan menambah bukti empiris mengenai dampak penggunaan *embedding* INDOBERTWEET sebagai instrumen evaluasi semantik untuk meningkatkan kualitas hasil *Gibbs Sampling Dirichlet Mixture Model* (GSDMM) dalam pemodelan topik teks pendek.

Secara praktis, hasil penelitian ini dapat memberikan rekomendasi yang relevan dan aplikatif untuk analisis opini pelanggan di media sosial berbahasa Indonesia. Dengan memahami topik-topik utama yang muncul dari opini pelanggan, perusahaan dapat memperoleh wawasan yang berharga untuk meningkatkan kualitas layanan berdasarkan kebutuhan dan keluhan pelanggan yang teridentifikasi dari data media sosial.