

BAB I

PENDAHULUAN

1.1 Latar Belakang

Di era digital yang semakin terintegrasi, media sosial telah menjadi salah satu platform bagi masyarakat Indonesia untuk menyuarakan opini. Di Indonesia, ekosistem digital semakin massif dan menjadikan media sosial sebagai kanal ekspresi opini publik. Laporan Digital 2025 Indonesia mencatat 143 juta identitas pengguna media sosial, jumlah itu setara dengan 50,2% dari total penduduk Indonesia per-Januari 2025, dengan potensi jangkauan Instagram sekitar 103 juta pengguna dari basis pengguna internet menunjukkan besarnya audiens percakapan kebijakan di *platform* media sosial seperti Instagram [1]. Survei APJII 2025 mengestimasi penetrasi internet sekitar 229 juta pengguna, menegaskan luasnya basis partisipasi warganet Indonesia dalam diskusi kebijakan publik [2]. Pada level konsumsi informasi, temuan *Digital News Report 2025* menandai peran sentral media sosial dalam akses berita, terutama di kelompok usia muda, yang relevan untuk memantau opini. Selain itu, data dari Meltwater 2024 menunjukkan bahwa Indonesia memiliki 85,3% pengguna aktif Instagram, di mana komentar pada postingan terkait kebijakan publik sering kali mencapai ribuan interaksi per hari, mencerminkan tren peningkatan keterlibatan warga dalam interaksi sosial yang terjadi di ruang digital [3].

Meluasnya penggunaan media sosial juga memunculkan tantangan serius seperti disinformasi, polarisasi, dan lemahnya moderasi platform. Pemerintah Indonesia dalam tahun 2025, beberapa kali menekan platform besar seperti TikTok dan Instagram untuk memperkuat moderasi menyusul konten yang memicu keresahan publik menunjukkan sensitifnya wacana kebijakan saat momentum politik/ekonomi memanas [4]. Di sisi lain percakapan di Instagram sering kali campur Indonesia-Inggris dan sarat slang atau variasi ejaan, memperberat analisis otomatis jika tidak ditangani dengan teknik yang tepat [5]. Hal ini mengimplikasikan perlunya pendekatan analisis data yang lebih canggih untuk mengungkap pola opini tersembunyi dan mencegah eskalasi konflik sosial. pemetaan opini berbasis komentar media sosial menyediakan sinyal dini respons Masyarakat atas kebijakan dan kinerja institusi, ketersediaan data komentar

Instagram yang *real-time* dan kaya konteks meski pendek menjadikannya sumber penting untuk menilai penerimaan dan risiko komunikasi kebijakan [6].

Pada masa kepemimpinan menteri keuangan bapak Purbaya Yudhi Sadewa, kebijakan dan tata kelola Direktorat Jenderal Bea dan Cukai menjadi salah satu isu menonjol dalam wacana publik. Dalam beberapa rapat dan pernyataan resmi, bapak Purbaya memberikan tenggat waktu sekitar satu tahun kepada Bea Cukai untuk memperbaiki kinerja, terutama terkait praktik penyelundupan, permainan klasifikasi HS code dan pungutan liar. Jika perbaikan tidak tercapai, ia membuka opsi pembekuan instansi dan merumahkan pegawai [25]. Menteri keuangan baru, bapak Purbaya menegaskan bahwa secara realistis tidak mungkin mencapai kondisi tanpa kecurangan sama sekali dalam pelayanan dan pengawasan ekspor dan impor yang dilakukan Direktorat Jenderal Bea dan Cukai, namun pemerintah menargetkan penurunan drastis praktik penipuan atau kejahatan lain melalui penguatan pengawasan dan perbaikan sistem pelayanan [26].

Penelitian terdahulu tentang pemodelan opini publik di media sosial telah banyak menggunakan metode *topic modeling* seperti *Latent Dirichlet Allocation* (LDA), namun masih terdapat kesenjangan dalam penanganan teks pendek dan dinamis seperti Komentar Instagram. Sebuah tinjauan oleh penelitian terdahulu menemukan bahwa LDA sering kali kurang efektif [7]. Meskipun *Top2Vec* telah diperkenalkan sebagai alternatif berbasis embedding sejak 2020, penerapannya pada konteks Indonesia masih terbatas, di mana studi seperti Angelov menyoroti kelemahan dalam integrasi reduksi dimensi UMAP untuk data multilingual yang kaya akan kode campur seperti bahasa Indonesia-Inggris di Instagram.

Penelitian sebelumnya belum banyak yang menganalisis Komentar Instagram bertema kebijakan keuangan menggunakan *Top2Vec* secara otomatis dan terintegrasi dari awal hingga akhir (end-to-end). Banyak studi masih berfokus ke klasifikasi sentimen atau platform lain seperti Twitter/Youtube. Kesenjangan tersebut menekankan pentingnya penelitian ini yang secara spesifik menerapkan *Top2Vec* untuk memodelkan opini terhadap kebijakan Menteri Keuangan bapak Purbaya, mengatasi keterbatasan metode sebelumnya melalui pendekatan *unsupervised* yang fleksibel. Dengan memanfaatkan *embedding vektor* untuk representasi semantik, *Top2Vec* memungkinkan ekstraksi topik tanpa pra-definisi

jumlah klaster, Penelitian ini akan mengisi gap dengan fokus pada *preprocessing* teks Instagram seperti *removal*, slang dan normalisasi, diikuti reduksi dimensi UMAP dan *clustering* HDBSCAN, untuk menghasilkan topik opini yang lebih akurat dan *interpretable*.

Penelitian ini menjadi semakin relevan di tengah kebijakan yang diterapkan oleh Menteri Keuangan baru tentang bea cukai, di mana opini publik di media sosial yang sangat melimpah dan dapat menjadi acuan penelitian. Analisis dari Jurnal terdahulu menunjukkan bahwa jaringan Twitter tentang isu Kementerian Keuangan sering kali membentuk kelompok yang menutup diri dari opini lain atau disebut *echo chamber* yang mempercepat penyebaran misinformasi, dengan implikasi penurunan investasi asing akibat ketidakpastian kebijakan [8]. Dengan menerapkan metode Top2Vec, penelitian ini menilai kualitas hasil pemodelan topik menggunakan *coherence score*, *information gain*, *topic diversity*, serta proporsi *outlier/noise*. Penelitian ini juga memberikan tema topik secara ringkas dengan merujuk pada top words dan contoh komentar representatif pada tiap *id topic*, sehingga pembaca dapat memahami gambaran umum tema percakapan yang muncul pada Komentar Instagram terkait kebijakan Menteri Keuangan tentang bea cukai. Dari sisi akademik, penelitian ini menambah referensi penerapan natural language processing (NLP) untuk pemodelan topik pada data media sosial yang cenderung informal dan bervariasi. Hasil pemodelan dapat menjadi masukan awal bagi pihak terkait, seperti Kementerian Keuangan, untuk mengenali tema yang dominan dibahas publik, misalnya pajak/biaya, rokok ilegal, atau pungli, sebagai dasar penyusunan komunikasi publik yang lebih tepat sasaran berdasarkan isu yang benar-benar muncul pada data. Secara keseluruhan, penelitian ini diharapkan memperkaya kajian pemodelan topik dan analisis opini publik pada isu kebijakan ekonomi di Indonesia.

1.2 Perumusan Masalah

1.2.1 Permasalahan

Terjadi peningkatan jumlah volume Komentar Instagram terkait kebijakan Menteri Keuangan baru tentang Bea Cukai pasca perubahan kabinet 2025 menyulitkan analisis manual karena data bersifat teks pendek, informal, banyak mengandung slang/variasi ejaan, serta sering terjadi campuran bahasa Indonesia

dan Inggris. Karakteristik tersebut meningkatkan potensi noise dan membuat pemetaan opini publik menjadi kurang konsisten apabila tidak ditangani dengan pendekatan yang mampu menangkap makna semantik pada komentar. Oleh karena itu, diperlukan metode pemodelan topik yang dapat mengelompokkan opini secara otomatis berdasarkan kemiripan semantik, tanpa perlu menentukan jumlah topik di awal, serta menghasilkan topik yang dapat dievaluasi secara kuantitatif dan dipahami secara kualitatif.

1.2.2 Metode Usulan

Penelitian ini mengusulkan penggunaan metode *Top2Vec* sebagai pendekatan utama untuk memodelkan opini publik, yang mengintegrasikan *embedding vektor*, reduksi dimensi dengan *Uniform Manifold Approximation and Projection* (UMAP), dan *clustering Vector* berbasis densitas menggunakan *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN). Metode ini bersifat *unsupervised*, sehingga tidak memerlukan penentuan jumlah topik diawal. Proses dimulai dari pengumpulan data melalui *Extensions Instagram Coment* Komentar Instagram, diikuti *preprocessing* (*case folding*, *cleaning*, normalisasi, tokenisasi, *stemming*), pengimplementasian *Top2vec*, serta evaluasi melalui *coherence score*, *topic diversity*, *information gain*, *outlier* dan visualisasi (UMAP *Scatter*, *Bar chart top words*, *word cloud*).

1.2.3 Pertanyaan Penelitian

Berdasarkan permasalahan di atas, pertanyaan penelitian difokuskan pada dua aspek utama:

1. Bagaimana karakteristik topik yang dihasilkan *Top2Vec* ditinjau dari *top words*, *topic size*, serta komentar representatif pada setiap *topic_id*?
2. Bagaimana pengaruh penggunaan UMAP pada skenario 1 dan skenario 2, ditinjau dari metrik evaluasi *coherence score*, *topic diversity*, *information gain* dan proporsi *outlier/noise*?

1.3 Tujuan & Manfaat

1.3.1 Tujuan Penelitian

Tujuan penelitian ini disesuaikan dengan pertanyaan penelitian untuk memastikan fokus yang terarah:

1. Mendeskripsikan karakteristik topik yang dihasilkan metode *Top2Vec* pada Komentar Instagram tentang bea cukai dengan meninjau top words, topic size, serta komentar representatif pada setiap topic_id.
2. Menganalisis pengaruh penggunaan UMAP pada Permodelan topik dengan membandingkan Skenario 1 Default *Top2Vec* dan Skenario 2 *Top2Vec* tanpa menggunakan UMAP, berdasarkan metrik evaluasi coherence score, topic diversity, information gain, serta proporsi outlier/noise.

1.3.2 Manfaat

Berdasarkan rumusan masalah yang ada, dapat ditentukan manfaat dari penelitian ini diantaranya adalah sebagai berikut:

1. Menambah referensi penerapan metode *Top2Vec* dalam pemodelan topik opini publik pada Komentar Instagram yang bersifat teks pendek dan informal.
2. Memberikan gambaran perbandingan hasil pemodelan *Top2Vec* pada dua skenario, dengan UMAP dan tanpa UMAP sehingga dapat menjadi rujukan dalam menentukan pemodelan topik yang lebih sesuai untuk data komentar media sosial.
3. Menyediakan evaluasi kualitas topik yang lebih terukur melalui *coherence score*, *topic diversity*, *information gain*, serta proporsi *outlier/noise*, sehingga mendukung pembahasan topik secara lebih objektif.
4. Menghasilkan pemetaan tema topik pada Komentar Instagram terkait kebijakan Bea Cukai, yang dapat dimanfaatkan sebagai masukan awal bagi instansi terkait untuk memahami fokus pembahasan publik.

1.4 Batasan-batasan

Penelitian ini dibatasi pada:

1. Data: Terbatas pada Komentar Instagram dari postingan akun yang membahas kebijakan Menteri Keuangan baru yakni bapak Purbaya Yudhi Sadewa. Pengambilan data dilakukan hingga menjelang pelaksanaan sidang skripsi. Data yang digunakan berbahasa Indonesia atau campuran Indonesia dan inggris, serta tidak mencakup analisis non-teks seperti gambar atau video.

2. Metodologi: Penelitian difokuskan pada penggunaan metode *Top2Vec* dengan tahapan *preprocessing* teks. Evaluasi model dibatasi pada metrik evaluasi utama, yaitu *coherence score*, *topic diversity*, *information gain* dan *outlier/noise*.
3. Kontekstual: Penelitian ini tidak membahas aspek etika data, pengaruh akun bot, maupun implikasi hukum. Populasi yang dianalisis dibatasi pada opini publik di platform Instagram.

1.5 Sistematika Penulisan Skripsi

Skripsi ini disusun secara sistematis sebagai berikut:

Bab I Pendahuluan

Meliputi latar belakang, perumusan masalah, tujuan dan manfaat, batasan, serta sistematika proposal.

Bab II Tinjauan Pustaka

Membahas dasar teori seperti opini publik dan media sosial, metode *Top2Vec*, *text mining*, *Natural Language Processing (NLP)*, semantik, *word embedding*, *Uniform Manifold Approximation and Projection (UMAP)*, *Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)*, serta metrik evaluasi seperti *coherence score*, *topic diversity*, dan *information gain*. Selain itu, bab ini juga memuat penelitian-penelitian terdahulu yang relevan sebagai dasar pengembangan penelitian ini.

Bab III Metodologi

Menjelaskan tahap penelitian yang dilakukan, mulai dari studi literatur, pengumpulan data, tahapan *preprocessing* teks, proses pemodelan *Top2Vec* meliputi *embedding*, reduksi dimensi, dan klasterisasi, hingga evaluasi model dan visualisasi hasil. Bab ini juga mencakup skenario uji coba untuk menilai pengaruh penggunaan algoritma *UMAP* terhadap kinerja *Top2Vec*.

Bab IV Hasil Dan Pembahasan

Penyajian dimulai dari data penelitian, jumlah data akhir setelah proses pembersihan dan karakteristik umum dataset. Selanjutnya dipaparkan hasil pemodelan *Top2Vec* pada masing-masing skenario pengujian, baik dengan maupun tanpa penggunaan *UMAP*, meliputi jumlah

topik yang terbentuk, distribusi dokumen pada setiap topik, serta jumlah noise atau outlier yang dihasilkan. Hasil evaluasi model ditampilkan menggunakan metrik coherence score, topic diversity, dan information gain untuk membandingkan kualitas topik yang terbentuk pada setiap skenario. Visualisasi hasil seperti grafik distribusi topik, tabel top words, *UMAP scatter*, *information gain heatmap* dan *word cloud*.

Bab V Penutup

Kesimpulan yang dirumuskan berdasarkan hasil analisis, disusun secara ringkas dan sistematis dengan menjawab rumusan masalah penelitian, khususnya terkait pengaruh penggunaan algoritma UMAP terhadap performa model Top2Vec dalam membentuk topik.

Daftar Pustaka

Daftar pustaka memuat seluruh referensi yang digunakan dalam penyusunan skripsi, baik berupa buku, jurnal ilmiah, prosiding, maupun sumber daring yang relevan. Penulisan daftar pustaka disusun secara sistematis sesuai dengan gaya sitasi.