BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Analisis opini pelanggan di media sosial merupakan bagian dari kajian yang lebih luas dalam bidang NLP (*Natural Language Processing*) dan *text mining*, yang berfokus pada bagaimana informasi tidak terstruktur dari pelanggan dapat diolah menjadi wawasan yang bermakna. Pemodelan topik merupakan metode pembelajaran mesin tanpa pengawasan (*unsupervised learning*) yang didasarkan pada statistik data untuk menemukan "topik" yang menggambarkan kumpulan dokumen (Lossio-Ventura dkk., 2021). Dengan menggunakan petunjuk kontekstual, pemodelan topik dapat menghubungkan kata-kata yang memiliki makna serupa dan membedakan kata-kata yang memiliki banyak makna (Hajjem dan Latiri, 2017). Pemodelan topik juga dapat dianggap sebagai metodologi untuk menyajikan volume data yang besar yang dihasilkan akibat perkembangan teknologi komputer dan web, ke dalam bentuk yang mudah dipahami dan juga untuk mengungkap informasi tersembunyi, fitur yang menonjol, atau variabel laten dari data (Kherwa dan Bansal, 2020).

Pemodelan topik adalah metode untuk menganalisis dokumen dengan masukan berupa kumpulan dokumen mentah, selanjutnya sistem mengidentifikasi dan mengelompokkan pola kata yang berulang menjadi topik atau tema. Setelah dokumen dimasukkan, algoritma pemodelan topik mengembalikan serangkaian tema/topik yang berkaitan dengan dokumen-dokumen tersebut. Studi-studi sebelumnya menunjukkan bahwa metode pemodelan topik ini mampu diterapkan dalam berbagai kasus, penelitian oleh Widiantoro dkk (2023) mengkaji pemanfaatan pemodelan topik untuk mengekstraksi fitur laten dari komentar pengguna aplikasi Financial Technology (FinTech) di Indonesia menggunakan algoritma *Latent Dirichlet Allocation* (LDA). Studi ini bertujuan mengidentifikasi isu dan kebutuhan utama pengguna berdasarkan komentar yang banyak dan bervariasi, untuk mendukung pengembangan fitur aplikasi FinTech melalui integrasi dengan kerangka User-Centered Design (UCeD). Penelitian menggunakan

dataset besar berupa lebih dari 52.000 komentar pengguna aplikasi P2P Lending yang diambil dari Google Play Store. Komentar tersebut melalui proses pra-pemrosesan yang meliputi normalisasi teks, penghilangan karakter khusus, stemming menggunakan alat Sastrawi, dan penyusunan matriks term-document untuk diolah oleh LDA. Hasil analisis LDA mengungkap sepuluh topik utama yang memiliki koherensi rata-rata sebesar 0.564, menunjukkan kualitas topik yang cukup baik dan bermakna secara semantik. Topik-topik tersebut kemudian dikelompokkan menjadi tiga kategori besar: perbaikan sistem aplikasi, layanan yang tidak sesuai dengan harapan pengguna, dan aspek kepuasan layanan. Evaluasi kualitas topik juga melibatkan diskusi dengan para ahli untuk memastikan relevansi dan interpretasi yang tepat terhadap setiap topik. Penelitian ini mengintegrasikan hasil LDA ke dalam proses UCeD, sehingga otomatisasi evaluasi, abstraksi, dan pengelompokan kebutuhan pengguna dapat dilakukan secara lebih efisien dan sistematis dalam pengembangan aplikasi FinTech. Kelebihan penelitian ini terletak pada penggunaan dataset yang besar dan nyata dari pengguna FinTech di Indonesia, serta integrasi yang inovatif antara teknik *topic modeling* dengan kerangka desain yang berpusat pada pengguna, yang memungkinkan peningkatan proses pengembangan fitur berbasis kebutuhan nyata pengguna. Selain itu, penggunaan metode koherensi topik sebagai metrik evaluasi memberikan validitas terhadap kualitas model. Namun, penelitian ini memiliki keterbatasan, antara lain ketergantungan pada penilaian manual dalam pelabelan topik dan evaluasi interpretasi, serta fokus yang terbatas pada LDA tanpa eksplorasi metode *topic modeling* lain atau berbasis *embedding*. Relevansi penelitian ini sangat tinggi dalam mengkaji topik modeling untuk ekstraksi fitur dari data teks pengguna, khususnya di bidang layanan digital atau aplikasi mobile. Pendekatan integrasi LDA dengan UCeD dapat menjadi referensi penting untuk memperbaiki metodologi pengolahan dan analisis komentar atau teks pendek, sekaligus membantu dalam pengambilan keputusan berbasis data yang mendalam terkait kebutuhan pengguna (Widiantoro dkk., 2023).

Penelitian oleh Abdelmotaleb dkk (2023) mengembangkan metode evaluasi dan *tuning hyperparameter* untuk model *topic modeling* GSDMM yang khusus

dirancang untuk teks pendek seperti komentar pelanggan di industri telekomunikasi. Dengan menggunakan representasi kata berbasis *embedding* FastText, mereka mengukur variasi dalam klaster dan jarak antar klaster untuk membentuk fungsi *cost* yang mengoptimalkan nilai hyperparameter α dan β. Menggunakan metrik *Within-Cluster Variation (WCV)* dan *Between-Cluster Variation (BCV)* berbasis *cosine distance*. Eksperimen pada dataset komentar pelanggan dari TrustPilot dan BT Community menunjukkan bahwa nilai optimal hyperparameter α dan β sebagian besar berada di rentang 0.5 hingga 0.9, berbeda signifikan dari nilai default 0.1 yang umum digunakan, yang menghasilkan cluster topik yang lebih representatif dan seimbang dari segi distribusi dokumen. Topik-topik yang ditemukan relevan dengan isu utama dalam layanan telekomunikasi seperti komunikasi, konektivitas internet, layanan pelanggan, dan kecepatan jaringan. Metode ini memberikan kontribusi penting dalam memperbaiki evaluasi dan tuning GSDMM secara sistematis, memungkinkan pemilihan parameter yang lebih tepat untuk hasil clustering yang berkualitas. Namun, penelitian ini terbatas pada domain telekomunikasi dengan dataset relatif kecil dan belum melakukan perbandingan dengan model-topic *embedding* modern lainnya. Relevansi penelitian ini sangat tinggi bagi studi yang menggunakan GSDMM atau model clustering teks pendek lainnya, karena menawarkan pendekatan evaluasi dan optimasi yang dapat meningkatkan performa model dalam mengelola data teks pendek seperti tweet atau komentar singkat. (Abdelmotaleb dkk., 2023).

Penelitian yang dilakukan oleh Khadija dan Nurharjadmo (2024) bertujuan untuk meningkatkan kualitas analisis keluhan pelanggan berbahasa Indonesia menggunakan kombinasi *Latent Dirichlet Allocation* (LDA) dan *Bidirectional Encoder Representations from Transformers* (BERT) untuk mengeksplorasi topik-topik utama dalam data keluhan pelanggan yang bersumber dari Twitter dan Instagram, dataset didapatkan dari DomaiNesia. Pendekatan ini dilatarbelakangi oleh keterbatasan LDA dalam mengelola teks pendek dan informal yang banyak ditemukan pada media sosial. Oleh karena itu, model BERT digunakan untuk memperkaya representasi kata dalam topik yang dibentuk, dengan harapan dapat meningkatkan kualitas semantik dan struktur dari hasil topik modeling. Penelitian

ini menerapkan metode pemodelan topik LDA dikombinasikan dengan dua jenis model BERT multibahasa, yaitu *Multilingual Cased* dan *Uncased*. Diawali dengan tahap pra-pemrosesan teks berupa normalisasi, penghapusan *stopwords*, dan stemming. Evaluasi terhadap hasil topik dilakukan dengan menggunakan dua metrik utama, yaitu coherence score untuk mengukur keterkaitan semantik antar kata dalam satu topik, serta silhouette score untuk menilai kejelasan pemisahan antar cluster topik. Hasil menunjukkan bahwa kombinasi LDA dan BERT Multilingual Cased menghasilkan *coherence score* tertinggi sebesar 0,388, sementara LDA dan BERT *Multilingual Uncased* mencapai *silhouette score* tertinggi sebesar 0,616. Temuan ini memperkuat argumen bahwa integrasi *embedding* kontekstual dari BERT mampu meningkatkan performa LDA dalam menangani data teks pendek berbahasa Indonesia, dibandingkan dengan metode konvensional seperti LDA dengan TF-IDF. Kelebihan utama dari studi ini terletak pada pendekatan hibrida yang efektif dalam mengatasi kelemahan semantik LDA, khususnya dalam konteks data singkat dan informal. Namun demikian, beberapa keterbatasan masih ditemukan, antara lain ukuran data yang terbatas dan kemunculan kata-kata tidak relevan akibat penggunaan BERT multibahasa yang tidak terlatih secara khusus pada korpus Indonesia. Kendati demikian, penelitian ini memiliki relevansi yang sangat tinggi terhadap penelitian ini, yang juga berfokus pada *topic modeling* terhadap data teks pendek. Penggunaan model BERT untuk memperkuat struktur semantik dalam proses ekstraksi topik dapat menjadi alternatif metodologis yang penting, terutama dalam menghadapi tantangan data ringkas seperti tweet atau komentar publik. Dengan demikian, pendekatan LsDA-BERT dapat dijadikan salah satu acuan dalam perancangan model analisis topik yang lebih representatif, interpretable, dan semantik-responsif dalam penelitian ini (Khadija dan Nurharjadmo, 2024).

Penelitian yang dilakukan oleh Garcia dan Berton (2021) bertujuan untuk mengidentifikasi topik dominan serta menganalisis sentimen dalam percakapan publik mengenai pandemi COVID-19 melalui platform Twitter. Penelitian ini membandingkan persepsi publik dari dua negara dengan kasus COVID-19 tertinggi, yaitu Amerika Serikat dan Brasil. Data yang digunakan terdiri dari sekitar

3,3 juta tweet dalam bahasa Inggris dan 3,1 juta tweet dalam bahasa Portugis yang dikumpulkan dalam kurun waktu empat bulan (April–Agustus 2020). Metodologi penelitian ini mengintegrasikan dua pendekatan yaitu pemodelan topik menggunakan GSDMM dan analisis sentimen menggunakan algoritma berbasis leksikon dan *embedding* semantik seperti CrystalFeel, SBERT, serta Multilingual Universal Sentence Encoder (mUSE). Pemilihan metode GSDMM didasarkan pada kemampuannya yang efektif dalam menangani teks pendek seperti tweet, yang sering kali tidak dapat ditangani secara optimal oleh metode konvensional seperti LDA. Hasil penelitian ini menemukan bahwa topik-topik utama di kedua negara relatif mirip, dengan isu seperti dampak ekonomi, statistik kasus, perawatan COVID-19, kebijakan politik, dan langkah-langkah pencegahan penyebaran virus yang mendominasi diskusi publik. Selain itu, analisis sentimen menunjukkan bahwa sebagian besar percakapan bersifat negatif, dengan dominasi emosi seperti ketakutan (fear) dan kemarahan (anger), terutama terkait kebijakan pemerintah dan data statistik COVID-19. Kelebihan penelitian ini terletak pada penggunaan dataset yang sangat besar, analisis yang mencakup dua bahasa, serta pendekatan kombinasi metode yang memungkinkan analisis yang lebih mendalam terhadap topik maupun sentimen. Di sisi lain, kelemahan yang tercatat adalah ketergantungan metode sentimen pada anotasi manual serta keterbatasan generalisasi karena fokusnya hanya pada data dari dua negara dengan kondisi spesifik. Penelitian Garcia dan Berton (2021) memiliki relevansi tinggi dengan penelitian ini, terutama dalam aspek metodologis. Pemanfaatan GSDMM sebagai model *topic modeling* dan penggunaan *embedding* semantik modern seperti SBERT dapat memberikan dasar metodologis yang kuat dalam menangani data tweet berbahasa Indonesia, yang juga bersifat singkat dan informal. Selain itu, pendekatan kombinasi analisis topik dan sentimen dapat menjadi inspirasi dalam merancang model yang lebih komprehensif, khususnya jika penelitian berorientasi pada eksplorasi isu publik melalui data media sosial.

Penelitian oleh Jang dkk. (2021) bertujuan untuk mengidentifikasi isu-isu utama serta persepsi publik terkait pandemi COVID-19 dengan memanfaatkan data media sosial twitter, dari Kanada dan Amerika Serikat. Studi ini menggabungkan

teknik pemodelan topik dan analisis sentimen berbasis aspek (Aspect-Based Sentiment Analysis/ABSA) untuk mengeksplorasi bagaimana masyarakat bereaksi terhadap kebijakan kesehatan publik dan isu-isu yang muncul selama pandemi. Fokus utama penelitian adalah untuk memahami opini publik secara mendalam melalui kolaborasi langsung dengan pakar kesehatan publik dalam interpretasi data. Metode kombinasi dari *Latent Dirichlet Allocation* (LDA) digunakan untuk pemodelan topik dan ABSA yang bersifat spesifik domain. Data penelitian terdiri dari sekitar 319.524 tweet berbahasa Inggris yang dikumpulkan dari Kanada (sekitar 25.595 tweet) dan Amerika Serikat (sekitar 293.929 tweet). Dalam proses ABSA, penelitian ini melibatkan pakar domain untuk menentukan aspek-aspek penting yang relevan dengan COVID-19, seperti "physical distancing," "masks," dan "misinformation," serta istilah opini khusus yang mencerminkan sentimen positif atau negatif terhadap aspek-aspek tersebut. Pendekatan ini memungkinkan analisis yang mendalam dan spesifik terhadap isu-isu publik yang paling relevan dalam pandemik. Hasil penelitian menunjukkan bahwa topik-topik yang muncul dari data Twitter sangat terkait dengan kebijakan intervensi kesehatan publik, seperti pembatasan perjalanan udara, pembatasan perbatasan regional, kebijakan tinggal di rumah, pemakaian masker, serta informasi terkait jumlah tes dan kasus COVID-19. Analisis sentimen berbasis aspek menunjukkan dominasi sentimen negatif terhadap pandemi secara umum, terutama terkait isu rasisme anti-Asia dan penyebaran informasi yang salah. Di sisi lain, sentimen positif dominan dalam aspek-aspek seperti physical distancing. Penelitian juga menemukan bahwa meskipun pola topik yang dibahas di Kanada dan Amerika Serikat memiliki kesamaan, terdapat perbedaan spesifik dalam tingkat prevalensi topik tertentu antara kedua negara. Penelitian ini relevan dengan penelitian yang sedang dilakukan, khususnya jika fokus penelitian adalah penggunaan *topic modeling* dan analisis sentimen berbasis aspek pada data media sosial untuk mengidentifikasi isu public (Jang dkk., 2021).

Metode awal pemodelan topik, seperti Latent Semantic Analysis (LSA) (Landauer Thomas dan Foltz Peter, 1998), mengandalkan teknik faktorisasi matriks, tetapi memiliki keterbatasan dalam menangani polisemi dan sinonim.

Polisemi adalah kata yang memiliki lebih dari satu makna dan sinonim adalah kata-kata yang memiliki arti sama. Salah satu metode yang paling banyak digunakan dalam pemodelan topik adalah *Latent Dirichlet Allocation* (LDA) yang dikembangkan oleh Blei dkk. (2003), yang memodelkan dokumen sebagai campuran topik, dan setiap topik sebagai distribusi kata. LDA memberikan kerangka kerja yang lebih kuat dibandingkan dengan pendekatan konvensional seperti *bag-of-words* (BoW) atau TF-IDF (Egger dan Yu, 2022; Odden dkk., 2024; Thakral dkk., 2024; Walsh dkk., 2024). LDA memodelkan dokumen sebagai campuran topik, di mana setiap topik direpresentasikan sebagai distribusi kata-kata. Pendekatan ini memungkinkan penemuan topik laten bahkan di tengah data yang berisik atau ambigu, sehingga sangat cocok untuk menganalisis teks media sosial. Metode LDA digunakan untuk menganalisis pengalaman pasien dengan modafinil, mengungkapkan wawasan tentang penggunaannya di berbagai kondisi kesehatan dan membandingkan persepsi pasien dengan bukti klinis yang ada (Walsh dkk., 2024). Namun, penelitian menunjukkan bahwa LDA kurang efektif ketika diaplikasikan pada teks pendek seperti tweet, karena ketergantungan tinggi terhadap informasi distribusi kata yang terbatas (Yin dan Wang, 2014; Qiang dkk., 2022).

Untuk mengatasi keterbatasan tersebut, dikembangkanlah *Gibbs Sampling Dirichlet Mixture Model* (GSDMM), yang dirancang khusus untuk menangani teks pendek. GSDMM mengasumsikan bahwa setiap dokumen hanya membahas satu topik utama, sehingga lebih sesuai dengan karakter tweet yang umumnya fokus pada satu isu per postingan (Yin dan Wang, 2014). Penelitian oleh Weisser dkk. (2023) menunjukkan bahwa GSDMM mampu menghasilkan pengelompokan topik yang lebih kohesif dibandingkan LDA pada data sosial media. Pada penelitian klustering layanan web, penerapkan metode GSDMM juga memiliki kinerja yang baik dibanding metode lainnya seperti LDA, LSA, CTM, HDP, (Agarwal dkk., 2020). Namun demikian, tantangan utama dalam penggunaan GSDMM adalah proses seleksi model terbaik, Oleh karena itu, ketergantungan pada evaluasi statistik semata tanpa validasi semantik (seperti yang ditawarkan oleh word embedding)

menjadi celah yang perlu diperbaiki untuk mendapatkan topik yang benar-benar berkualitas (De Santis dkk., 2024).

Diperlukanlah teknik *word embedding* untuk mengisi kelemahan tersebut. Pendekatan klasik seperti *bag-of-words* (BoW) dan term frequency-inverse document frequency (TF-IDF) telah digunakan secara luas karena kesederhanaannya. Namun, kedua metode ini hanya merepresentasikan kata berdasarkan frekuensi tanpa mempertimbangkan hubungan semantik maupun konteks kalimat, yang menyebabkan kehilangan makna penting, terutama pada teks pendek seperti tweet (Manning, 2009). Untuk mengatasi keterbatasan tersebut, dikembangkanlah pendekatan representasi berbasis *word embedding*, yang memetakan setiap kata ke dalam ruang vektor berdimensi tinggi berdasarkan pola kemunculan kata dalam korpus besar. Pendekatan ini memungkinkan model mengenali kata-kata yang memiliki makna serupa, meskipun secara literal berbeda. Dua model yang cukup populer dalam *embedding* statis adalah Word2Vec dan GloVe, yang berhasil menangkap kedekatan semantik antar kata dengan cara menghitung ko-occurrence atau konteks jendela (Mikolov dkk., 2013; Pennington dkk., 2014)

Pendekatan *embedding* statis juga masih memiliki keterbatasan dalam membedakan makna kata yang bergantung pada konteks. Untuk itu, hadir pendekatan baru berbasis transformer yang menghasilkan contextualized word *embedding*s, seperti yang diperkenalkan oleh BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin dkk., 2019). BERT memperhitungkan konteks di kedua arah sebelum dan sesudah kata dalam kalimat, sehingga dapat menangkap makna yang lebih dalam dan kontekstual. Pendekatan *word embedding* berbasis deep learning seperti Word2Vec, GloVe, dan BERT mulai digunakan untuk memperkaya representasi kata dalam vektor berdimensi tinggi. *Embedding* ini memungkinkan model memahami hubungan semantik antar kata dengan lebih akurat (Romero dkk., 2024; Si dkk., 2019). Dalam konteks bahasa Indonesia, IndoBERTweet merupakan model *embedding* yang dikembangkan secara khusus dengan pelatihan menggunakan 409 juta token dari tweet berbahasa Indonesia (Koto dkk., 2021), sehingga lebih sensitif terhadap gaya bahasa informal dan

struktur bahasa khas media sosial. Pemanfaatan IndoBERTweet sebagai metrik evaluasi eksternal (external evaluation metric) untuk memvalidasi kualitas topik keluaran GSDMM masih belum banyak dieksplorasi. Penelitian ini mengisi celah tersebut dengan menggunakan representasi vektor IndoBERTweet untuk mengukur kepadatan klaster (*cluster compactness*) dan pemisahan antar topik, sehingga pemilihan parameter model didasarkan pada kedekatan makna yang sesungguhnya, bukan sekadar probabilitas statistik. (Aksoy dkk., 2023; Hajek dkk., 2021).

Kajian literatur diatas merupakan sumber pustaka metode pemodelan topik dengan model LDA maupun GSDMM dan pendekatan *word embedding* sebagai pendekatan yang dapat meningkatkan performa analisis topik opini pelanggan di media sosial, baik dari segi akurasi topik maupun interpretabilitas hasil. Dengan dasar kajian tersebut, maka tujuan dilakukan penelitian ini berfokus pada pencarian hyperparameter optimal GSDMM dengan menggunakan metrik evaluasi semantik, yaitu Within-Cluster Variation (WCV), Between-Cluster Variation (BCV), dan Cost Function (CF). Pendekatan ini bertujuan untuk menjamin bahwa topik yang dihasilkan tidak hanya valid secara statistik, tetapi juga memiliki interpretabilitas tinggi bagi *provider* telekomunikasi dalam memahami keluhan dan kebutuhan pelanggan.

2.2 Dasar Teori

Dasar teori pada penelitian ini merupakan teori-teori relevan yang mendukung penelitian yang dilakukan. Dasar teori pada penelitian ini meliputi media sosial X, Pemodelan Topik, GSDMM, *Embedding Word*, IndoBERTweet, *Coherren Score, Perplexity* dan Metrik Evaluasi Semantik

2.2.1 Media Sosial X

X, sebelumnya dikenal sebagai Twitter, adalah salah satu platform media sosial yang paling banyak digunakan untuk berbagi informasi, berinteraksi, dan berdiskusi dalam skala global. Sejak diluncurkan pada tahun 2006, Twitter telah mengalami berbagai perkembangan hingga akhirnya mengalami *rebranding* menjadi X pada tahun 2023. Platform ini memungkinkan pengguna untuk mengunggah pesan singkat atau tweet yang awalnya dibatasi hingga 140 karakter dan kemudian diperluas menjadi 280 karakter. Dengan fitur seperti retweet,

mention (@), hashtag (#), dan like, X menjadi wadah utama bagi komunikasi publik dan penyebaran informasi secara real-time. Popularitasnya yang tinggi menjadikan X sebagai salah satu sumber data terbesar dalam studi media sosial, termasuk dalam penelitian mengenai opini pelanggan terhadap suatu produk atau layanan.

Media sosial memiliki peran strategis sebagai wadah utama untuk menyalurkan opini, keluhan, dan pengalaman pelanggan di mana saja dan kapan saja dalam konteks bisnis dan layanan pelanggan (Jha, 2019). X menjadi medium utama bagi konsumen untuk menyampaikan opini, baik dalam bentuk keluhan, pujian, saran, maupun diskusi terkait suatu layanan. Platform ini memiliki karakteristik unik dengan format teks pendek yang memungkinkan pengguna berbagi pengalaman dengan cepat dan mudah (Alipour dkk., 2023). Industri telekomunikasi di Indonesia, misalnya, sering kali menjadi subjek pembicaraan di X, di mana pelanggan memberikan komentar mengenai kualitas layanan, harga, kecepatan internet, hingga pengalaman mereka dengan operator tertentu. Salah satu keunggulan utama X sebagai sumber data opini pelanggan adalah kemampuannya menyediakan data dalam jumlah besar dan dalam waktu nyata. Tidak seperti survei konvensional yang memerlukan waktu dalam pengumpulan dan analisis, opini pelanggan di X dapat diperoleh secara instan, sehingga memungkinkan monitoring tren pelanggan dan analisis sentimen secara langsung.

Analisis opini pelanggan di X menghadapi berbagai tantangan walaupun memiliki kekayaan data. Format teks pendek dalam setiap tweet sering kali menyebabkan kurangnya konteks yang memadai. Dalam literatur *Natural Language Processing* (NLP), tweet diklasifikasikan sebagai teks pendek (*short text*) yang memiliki karakteristik sangat spesifik. Berbeda dengan dokumen panjang konvensional seperti artikel berita atau jurnal, teks pendek didefinisikan sebagai dokumen dengan panjang yang sangat terbatas, umumnya terdiri kurang dari 50 kata per dokumen atau dibatasi oleh jumlah karakter yang ketat (maksimal 280 karakter pada media sosial X). Keterbatasan jumlah kata ini menyebabkan fenomena kelangkaan data (*data sparsity*), di mana jumlah kata yang muncul dalam satu dokumen sangat sedikit dibandingkan dengan total kosa kata dalam korpus,

sehingga menyulitkan metode konvensional untuk menemukan pola kemunculan kata bersama (*co-occurrence*) yang kuat.

Selain masalah panjang teks, pengguna cenderung menggunakan bahasa informal, slang, emoji, dan singkatan, yang semakin memperumit pemrosesan data. Karakteristik lain dari X yang perlu diperhatikan adalah noise dalam data, di mana banyak tweet yang mengandung informasi tidak relevan, spam, atau iklan yang dapat mengganggu akurasi analisis. Tidak hanya itu, tren di X bersifat sangat dinamis, di mana topik pembicaraan dapat berubah dalam hitungan jam atau hari, sehingga model analisis yang diterapkan harus mampu beradaptasi dengan cepat terhadap perubahan pola komunikasi pengguna.

Media sosial X telah menjadi salah satu sumber data utama dalam berbagai studi dalam penelitian NLP dan analisis opini. Berbagai penelitian telah menggunakan tweet untuk analisis sentimen pelanggan, pemodelan topik, serta pemantauan krisis dan manajemen reputasi. Pendekatan seperti *Latent Dirichlet Allocation* (LDA) telah digunakan dalam pemodelan topik, tetapi sering mengalami kesulitan dalam menangani teks pendek karena terbatasnya jumlah kata dalam setiap tweet. Alternatif lain seperti *Gibbs Sampling Dirichlet Mixture Model* (GSDMM) telah terbukti lebih efektif dalam mengelompokkan teks pendek, namun masih memiliki keterbatasan dalam menangkap hubungan semantik antar kata. Oleh karena itu, penggunaan model *embedding* berbasis IndoBERTweet, yang dikembangkan secara khusus untuk bahasa Indonesia di media sosial, diharapkan dapat meningkatkan interpretabilitas dan kualitas pemodelan topik dalam analisis opini pelanggan di X.

Penelitian ini memilih X sebagai sumber data utama dalam analisis opini pelanggan terhadap layanan telekomunikasi di Indonesia, berdasarkan karakteristik dan tantangan yang telah dibahas. Dengan menerapkan GSDMM dengan IndoBERTweet sebagai evaluator semantik, penelitian ini bertujuan untuk meningkatkan efektivitas pemodelan topik pada teks pendek, sehingga dapat memberikan wawasan yang lebih akurat dan bermanfaat bagi *provider* telekomunikasi dalam meningkatkan kualitas layanan mereka.

2.2.2 Pemodelan Topik

Pemodelan topik adalah teknik statistik untuk menemukan topik laten dalam sekumpulan dokumen/sumber data. Analisis topik memanfaatkan *text mining* untuk menemukan topik semantik laten dalam badan teks (Alash dan Al-Sultany, 2020). Dokumen/sumber data biasanya berisi campuran topik dan setiap topik terdiri dari sekumpulan kata dan dapat memperoleh topik. Sekumpulan topik tersebut terdiri dari kata-kata yang sering muncul dan dapat memperoleh topik dengan menghubungkan kata-kata yang mempunyai arti serupa dan membedakan penggunaan kata-kata yang mempunyai arti ganda melalui penemuan kata-kata yang membantu untuk menentukan batasan antar subjek atau menemukan pola data yang dapat digunakan untuk mencapai suatu tujuan (Mohammed dan Al-Augby, 2020).

Analisis topik dapat diklasifikasikan menjadi 2 (dua) model, yaitu model probabilistik dan model non-probabilistik (Kherwa dan Bansal, 2020). Pendekatan probabilistik merupakan model yang generatif dan algoritma *unsupervised learning* di mana dokumen direpresentasikan sebagai campuran acak atas topik laten, di mana topik tertentu dicirikan oleh distribusi kata (Kalepalli dkk., 2020). Pertama kali diusulkan yaitu teknik LSA (Latent Semantic Analysis) (Landauer Thomas dan Foltz Peter, 1998) merupakan non-probabilistik model yang tujuan utamanya adalah menguraikan matriks istilah dalam dokumen dengan teknik dekomposisi nilai tunggal untuk mempelajari fitur laten. Teknik pemodelan topik LSA merupakan metode yang paling sederhana yang memberikan hasil yang lebih baik dibandingkan dengan representasi vektor, tetapi kelemahan utama dari metode ini adalah tidak dapat menemukan makna ganda dari suatu istilah. Karena representasi tingkat rendah dari layanan yang disediakan oleh LSA, metode ini tidak efisien, dan ada kekurangan penyematan yang dapat ditafsirkan. Dalam model ini, tidak ada latar belakang metode statistik yang kuat. Untuk mengatasi keterbatasan LSA, sebuah model baru diusulkan, yang merupakan varian probabilistik dari LSA dan memiliki dasar statistik yang dinamakan PLSA (Probabilistic Latent Semantic Analysis) (Hofmann, 1999). Model probabilistik dari PLSA juga disebut model aspek yang digunakan untuk menganalisis kemunculan bersama istilah dan

dokumen dalam data. Masalah utama dari PLSA adalah PLSA memiliki beberapa kelemahan, di antaranya adalah tidak dapat langsung memberikan probabilitas pada dokumen baru, sehingga kurang fleksibel dalam menangani data di luar kumpulan pelatihan. Selain itu, jumlah parameter dalam model bertambah secara linear seiring bertambahnya dokumen, yang dapat menyebabkan overfitting dan membuat model sulit diskalakan untuk dataset besar. PLSA juga tidak memiliki mekanisme regularisasi bawaan, sehingga lebih rentan terhadap kompleksitas berlebihan. Akibatnya, meskipun PLSA efektif dalam menemukan topik dalam teks, penggunaannya menjadi kurang praktis untuk dataset yang sangat besar atau yang terus berkembang.

Metode LDA (*Latent Dirichlet Allocation*) dipresentasikan oleh Blei, Ng, dan Jordan (2003), yang merupakan model generatif untuk pemodelan topik dalam kumpulan dokumen. LDA bekerja dengan menebak topik apa saja yang ada dalam dokumen, tanpa perlu diberi tahu sebelumnya. Model ini menggunakan prinsip probabilitas untuk menentukan seberapa besar kemungkinan sebuah kata muncul dalam suatu topik dan seberapa besar kemungkinan sebuah dokumen berisi topik tertentu. Dengan begitu, kita bisa mengelompokkan dokumen berdasarkan topik yang tersembunyi di dalamnya, meskipun kita tidak tahu topik-topik itu sejak awal Misalnya, jika ada artikel tentang sepak bola, kemungkinan besar topik di dalamnya berisi kata-kata seperti "gol," "pemain," dan "liga". Meskipun model LDA mengatasi kekurangan dari PLSA, model ini tidak dapat mengambil korelasi antar topik (Blei dkk., 2002).

Correlated Topic Model (CTM) adalah metode pemodelan topik probabilistik lain yang memperluas LDA dengan menemukan korelasi topik yang kaya (Blei dan Lafferty, 2006). Ide utama yang digunakan dalam model ini untuk menentukan korelasi adalah dengan mengganti proporsi topik dari metode LDA dasar dengan distribusi normal logistik. Perbedaan mendasar antara model LDA dan CTM adalah bahwa dalam LDA, diasumsikan bahwa topik-topik bersifat independen karena adanya dirichlet prior pada distribusi topik. CTM mengadopsi distribusi normal logistik atas proporsi topik, yang menentukan korelasi antar topik dengan menggunakan matriks kovarian (Aznag dkk., 2013) dan bobot topik

dihasilkan dari distribusi Gaussian. Meskipun representasi yang lebih kaya ini telah ditingkatkan, CTM tidak efisien untuk data yang luas. Selain itu, pemodelan korelasi berpasangan membebankan kompleksitas inferensi sebesar O(T3) di mana T adalah jumlah topik. Dalam teknik pemodelan topik, diperlukan penyetelan hiper-parameter untuk mengamati jumlah topik yang optimal. Hal ini menyebabkan beberapa masalah seperti waktu pelatihan yang tinggi dan efisiensi model yang tidak substansial. Dalam model-model ini, kelemahan lainnya adalah tidak adanya pembagian topik di antara dokumen-dokumen dalam kumpulan dataset. Sebagai contoh, jika kita perlu mengambil dokumen mengenai 'pendanaan universitas', maka topik-topik tersebut harus diambil untuk 'pendidikan' dan 'keuangan'.

Teknik pemodelan topik *Hierarchical Dirichlet Processing* (HDP) diperkenalkan utuk mengatasi masalah tersebut. Pada dasarnya, model ini adalah model Bayesian non parametrik yang menangkap jumlah topik yang optimal dengan sendirinya dan memberikan izin untuk mencampurkan komponen yang akan dibagi di antara dokumen (Teh dkk., 2005). Model HDP memiliki hirarki Proses Dirichlet, sehingga struktur seperti pohon memiliki keterbatasan yang membatasi fleksibilitas model.

Berbagai model pemodelan topik telah dikembangkan, mulai dari LSA hingga model probabilistik seperti LDA, CTM, dan HDP. Namun, model-model konvensional ini seringkali kurang efisien untuk teks pendek seperti tweet. Untuk mengatasi keterbatasan ini, model seperti Dirichlet Multinomial Mixture (DMM) dan khususnya *Gibbs Sampling Dirichlet Mixture Model* (GSDMM) dikembangkan dengan asumsi setiap dokumen hanya terkait dengan satu topik utama, menjadikannya lebih sesuai untuk karakteristik teks pendek (Yin dan Wang, 2014).

2.2.3 *Gibbs Sampling Dirichlet Mixture Model* (GSDMM)

Gibbs Sampling Dirichlet Mixture Model (GSDMM) adalah salah satu teknik pemodelan topik. Model ini dikembangkan oleh Yin dan Wang (2014), dirancang untuk mengatasi tantangan spesifik dalam pengelompokan teks pendek, dimana konten terbatas membuatnya sulit untuk mengidentifikasi topik menggunakan metode tradisional seperti *Latent Dirichlet Allocation* (LDA) (Lossio-Ventura dkk., 2021). Berbeda dengan LDA, yang mengasumsikan bahwa

setiap dokumen memiliki distribusi topik yang beragam, GSDMM menggunakan pendekatan berbasis clustering, di mana setiap dokumen ditugaskan ke satu topik secara eksklusif. Model ini sering dianalogikan dengan "*Movie Group Process*", yang menggambarkan proses pengelompokan dokumen seperti penonton bioskop yang harus memilih salah satu dari beberapa film yang tersedia, dengan kemungkinan berpindah berdasarkan preferensi yang berubah hingga penonton yang memiliki preferensi yang sama melihat film yang sama. Karena ketersebaran data dan masalah dimensi yang tinggi, teknik pemodelan topik tradisional seperti LDA, PLSA tidak sesuai untuk layanan dokumen, yang dijelaskan dalam bentuk teks pendek (Egger dan Yu, 2022; Qiang dkk., 2022). Gambaran Model GSDMM dapat dilihat pada Gambar 2.2 (Qiang dkk., 2022b).

Gambar 2.1 Gambaran Model GSDMM

Gambar 2.2 merupakan representasi grafis dari model *Gibbs Sampling Dirichlet Mixture Model* (GSDMM) dalam bentuk graphical model probabilistik. Model ini menggambarkan proses generatif dari dokumen berbasis asumsi bahwa setiap dokumen hanya berkaitan dengan satu topik utama, sebuah pendekatan yang lebih sesuai untuk data teks pendek seperti tweet, ulasan produk singkat, atau judul berita. Dalam model ini, simbol α merepresentasikan hiperparameter Dirichlet yang digunakan untuk menghasilkan distribusi topik per dokumen (θ). Meskipun dalam model tradisional seperti LDA, setiap dokumen memiliki distribusi topik (θ), pada GSDMM pendekatan ini disederhanakan sehingga setiap dokumen hanya dipetakan ke satu topik spesifik (z). Topik tersebut kemudian digunakan untuk menghasilkan kata-kata (w) dalam dokumen, berdasarkan distribusi kata dari topik tersebut (φ_k), yang berasal dari prior Dirichlet lainnya, yaitu β. Variabel n_d menunjukkan jumlah kata dalam dokumen ke-d, sedangkan kotak besar yang membungkus z dan w menunjukkan bahwa proses tersebut diulang untuk seluruh korpus dokumen D.

Kotak kecil di dalamnya menunjukkan bahwa setiap dokumen memiliki beberapa kata, dan setiap kata dihasilkan berdasarkan topik yang dipilih.

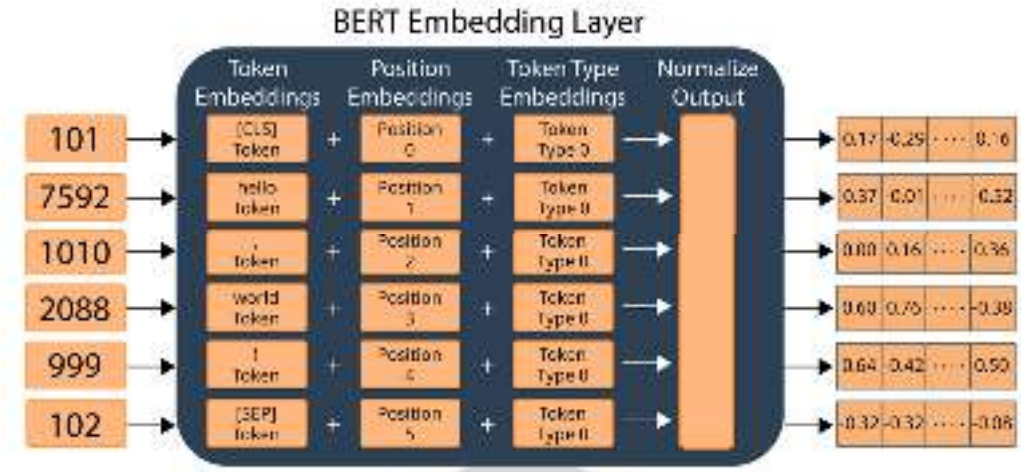
Berbeda dengan model *Latent Dirichlet Allocation* (LDA), yang mengasumsikan bahwa dokumen adalah campuran dari beberapa topik, GSDMM menggunakan pendekatan mixture model, yaitu satu dokumen satu topik. Pendekatan ini terbukti lebih efektif untuk menangani data teks pendek, yang secara alami memiliki cakupan semantik yang sempit dan kurang informasi distribusi kata. Model ini menggunakan teknik *Gibbs Sampling* untuk memperkirakan parameter distribusi topik dan distribusi kata dalam setiap topik, dengan efisiensi komputasi yang lebih baik serta hasil topik yang lebih kohesif dibandingkan metode konvensional dalam domain teks pendek. Dalam penerapannya, GSDMM bekerja dengan pendekatan *Dirichlet Process Mixture Model* dan menggunakan metode *Gibbs Sampling* untuk mengoptimalkan pengelompokan dokumen. Proses ini dimulai dengan inisialisasi awal, di mana setiap dokumen secara acak ditempatkan ke dalam salah satu dari K topik yang telah ditentukan. Selanjutnya, dalam iterasi *Gibbs Sampling*, dokumen akan dievaluasi untuk menentukan apakah lebih cocok berada dalam topik yang sama atau harus berpindah ke topik lain. Evaluasi ini didasarkan pada dua faktor utama, yaitu seberapa besar kecocokan kata dalam dokumen dengan kata-kata dominan dalam topik saat ini, serta seberapa besar ukuran topik tersebut dibandingkan dengan topik lainnya. Proses ini berlangsung secara iteratif hingga mencapai konvergensi, yaitu ketika tidak ada lagi perubahan signifikan dalam distribusi topik.

Dibandingkan dengan metode lain, GSDMM memiliki beberapa keunggulan utama. Pertama, model ini lebih efektif untuk teks pendek, karena mengasumsikan bahwa setiap dokumen hanya memiliki satu topik utama, yang sesuai dengan karakteristik tweet, komentar pelanggan, atau ulasan singkat. Kedua, metode ini lebih cepat dan efisien, karena menggunakan pendekatan berbasis clustering dengan *Gibbs Sampling* yang lebih ringan secara komputasi dibandingkan model probabilistik lain seperti LDA. Ketiga, GSDMM tidak memerlukan banyak hyperparameter, sehingga lebih mudah diimplementasikan tanpa perlu penyetelan parameter yang kompleks. Namun, GSDMM juga memiliki

beberapa keterbatasan, di antaranya sensitif terhadap nilai *hyperparameternya* yaitu jumlah optimal topik (K) alpha (α) dan beta (β), kurang efektif untuk dokumen panjang, serta keterbatasannya dalam memahami hubungan semantik antar kata karena masih mengandalkan pendekatan frekuensi kata *bag-of-words* (BoW).

2.2.4 *Embedding Word*

Dalam *Natural Language Processing* (NLP), representasi kata atau *word embedding* merupakan teknik yang digunakan untuk merepresentasikan kata-kata dalam bentuk vektor numerik dalam ruang berdimensi tinggi. Dimensi tinggi berarti kita menggunakan jauh lebih banyak dari sekadar 2 atau 3 angka untuk merepresentasikan setiap kata. Dalam teknik *word embedding* modern, jumlah dimensi ini seringkali mencapai ratusan (misalnya, 50, 100, 300, atau bahkan 768 atau lebih untuk model seperti BERT). Setiap kata diwakili oleh daftar panjang angka di mana setiap angka (dimensi) dalam vektor tersebut secara implisit menangkap suatu aspek dari makna atau konteks kata tersebut. Karena ada banyak dimensi, representasi ini dapat menangkap nuansa makna dan hubungan semantik yang kompleks antar kata. Misalnya, kata-kata yang memiliki makna serupa atau sering muncul dalam konteks yang sama akan memiliki vektor yang "mirip" atau "dekat" satu sama lain dalam ruang dimensi tinggi ini. Jadi, vektor untuk "raja" mungkin akan lebih dekat dengan vektor untuk "ratu" atau "istana" dibandingkan dengan vektor untuk "pisang" atau "mobil". Teknik ini memungkinkan pemodelan hubungan semantik antar kata berdasarkan konteks penggunaannya dalam suatu korpus teks. *Word embedding* menjadi solusi atas keterbatasan pendekatan tradisional seperti *bag-of-words* (BoW) dan *Term Frequency-Inverse Document Frequency* (TF-IDF), yang hanya mengandalkan frekuensi kata tanpa mempertimbangkan makna kontekstual. Ilustrasi bagaimana lapisan embedding BERT mengkonversi untuk string “hello, world” ke bentuk vector berdimensi tinggi (hingga 768) dapat dilihat pada Gambar 2.3 (Devlin dkk., 2019).



Gambar 2.2 Ilustrasi Konversi Kata ke Bentuk Vektor oleh BERT

Pendekatan *word embedding* yang lebih canggih berkembang dengan teknik berbasis distribusi semantik, seperti Word2Vec dan GloVe. Word2Vec, diperkenalkan oleh menggunakan pendekatan *Continuous Bag-of-words* (CBOW) dan Skip-Gram, yang bertujuan untuk memprediksi kata berdasarkan kata-kata sekitarnya (Mikolov dkk., 2013). Teknik ini memungkinkan kata-kata yang memiliki makna serupa untuk dikelompokkan lebih dekat dalam ruang vektor. Sementara itu, *Global Vectors for Word Representation* (GloVe) yang dikembangkan menggunakan pendekatan berbasis statistik yang menghitung frekuensi kemunculan kata dalam konteks tertentu untuk membangun representasi vektor kata yang lebih bermakna (Pennington dkk., 2014).

Teknik *embedding* statis seperti Word2Vec dan GloVe telah memiliki kemampuan menangkap hubungan semantik antar kata, namun metode ini masih memiliki keterbatasan dalam menangkap makna kontekstual. Kata-kata dengan makna ganda (polisemi) atau yang memiliki arti berbeda tergantung pada konteks penggunaannya tetap direpresentasikan dengan satu vektor tetap. Untuk mengatasi keterbatasan ini, dikembangkan model *embedding* berbasis transformer, seperti BERT (*Bidirectional Encoder Representations from Transformers*), yang mampu menangkap makna kata berdasarkan konteks di sekitarnya.

BERT, merupakan model deep learning berbasis transformer yang dapat memahami kata dalam dua arah (*bidirectional*) (Devlin dkk., 2019). Berbeda

dengan model sebelumnya yang membaca teks dari kiri ke kanan atau dari kanan ke kiri, BERT membaca teks secara simultan dalam kedua arah. Sehingga dapat memahami hubungan antar kata dalam konteks kalimat yang lebih kompleks, yang juga berakibat tingginya vector yang dihasilkan. Sebagai contoh model BERT-base menggunakan panjang vektor (atau dimensi embedding) sebesar 768 dan Model BERT-large menggunakan panjang vektor sebesar 1024. Model ini telah membuka jalan bagi pengembangan *embedding* kontekstual, yang lebih efektif dalam menangkap makna kata dalam berbagai situasi.

2.2.5 IndoBERTweet

IndoBERTweet adalah model *embedding* berbasis transformer yang dikembangkan secara khusus untuk menangani teks berbahasa Indonesia di media sosial. Model ini merupakan adaptasi dari BERT, tetapi dilatih menggunakan data dari media sosial Twitter (X), untuk menangkap pola bahasa yang unik dalam komunikasi digital. Salah satu tantangan utama dalam analisis teks media sosial adalah variasi bahasa yang sangat tinggi, termasuk penggunaan bahasa informal, singkatan, slang, serta campuran bahasa (code-mixing). Model NLP yang dilatih pada korpus teks formal sering kali kurang efektif dalam memahami teks media sosial karena perbedaan struktur bahasa dan gaya penulisan. IndoBERTweet dikembangkan dengan menggunakan dataset yang terdiri dari 409 juta token dari tweet berbahasa Indonesia, sehingga mampu menangkap nuansa bahasa yang digunakan dalam percakapan di media sosial dengan lebih baik dibandingkan model NLP konvensional (Koto dkk., 2021). Keunggulan utama IndoBERTweet dibandingkan *embedding* tradisional seperti Word2Vec dan GloVe adalah kemampuannya dalam menangkap makna kata berdasarkan konteks kalimat secara dinamis. Model ini menggunakan self-attention mechanism yang memungkinkan pemahaman kata dalam hubungan dengan kata lain di dalam kalimat. Dengan demikian, IndoBERTweet dapat mengatasi tantangan dalam analisis teks pendek, seperti polisemi dan perubahan makna kata berdasarkan konteksnya.

IndoBERTweet akan menjadi komponen krusial dalam penelitian ini, dengan keunggulannya dalam menangkap nuansa bahasa informal di media sosial Indonesia. Representasi vektor yang dihasilkan oleh IndoBERTweet diharapkan

dapat mengatasi keterbatasan semantik GSDMM yang berbasis frekuensi kata, sehingga mampu menghasilkan topik yang lebih akurat, koheren, dan mudah diinterpretasikan dari opini pelanggan.

2.2.6 *Coherren Score*

Coherence score adalah metrik evaluasi yang digunakan dalam pemodelan topik untuk mengukur kualitas dan interpretabilitas topik yang dihasilkan oleh model. *Coherence score* sangat penting dalam analisis topik karena topik yang dihasilkan oleh model harus dapat dipahami secara logis oleh manusia agar dapat digunakan untuk pengambilan keputusan. Metrik ini mengevaluasi seberapa baik kata-kata dalam suatu topik memiliki keterkaitan semantik satu sama lain, sehingga dapat mencerminkan konsistensi makna dalam topik yang dihasilkan (Abdelrazek dkk., 2023). Jika kata-kata dalam sebuah topik sering muncul bersamaan dalam berbagai dokumen, maka topik tersebut dianggap memiliki koherensi yang tinggi. Sebaliknya, jika kata-kata dalam satu topik tidak memiliki hubungan yang jelas atau jarang muncul bersama, maka topik tersebut dianggap kurang koheren (Chang dkk., 2009). Berbagai pendekatan telah dikembangkan untuk menghitung coherence score. Pendekatan yang paling banyak digunakan adalah *UMass Coherence* (Mimno dkk., 2011) dan *UCI Coherence* (Chang dkk., 2009) *Cv Coherence* (Röder dkk., 2015).

UMass Coherence mengukur hubungan antara pasangan kata dalam satu topik berdasarkan koeksistensi kata dalam dokumen. Dihitung berdasarkan frekuensi kemunculan bersama kata-kata dalam korpus dan sering digunakan untuk model berbasis probabilistik seperti LDA dan GSDMM. Matrik ini memiliki kekurangan dalam ketergantungannya terhadap frekuensi kemunculan kata dalam dokumen, sehingga lebih cocok untuk dokumen panjang daripada teks pendek seperti tweet. UMass biasanya menghasilkan skor negatif (semakin mendekati nol semakin baik). *UMass Coherence* menggunakan log conditional probability yang mengukur sejauh mana kata-kata dalam satu topik muncul bersamaan dalam dokumen. Meski relatif sederhana, pendekatan ini cukup efektif dalam menilai kedekatan kata dalam satu dokumen, meskipun korelasinya dengan penilaian

manusia tidak setinggi Cv atau NPMI Coherence (Korenčić dkk., 2021; Amaro & Bacao, 2024)

UCI Coherence menggunakan matrik berbasis *Pointwise Mutual Information* (PMI) untuk mengukur hubungan antara kata dalam topik berdasarkan frekuensi kemunculan bersama dalam sliding window tertentu. Lebih fleksibel dibandingkan UMass dan lebih cocok untuk menangani berbagai jenis teks, termasuk teks pendek seperti tweet. Metode ini efektif, namun dinilai lebih sensitif terhadap variasi frekuensi kata dan lebih rendah korelasinya dengan persepsi manusia dibandingkan metode Cv dan NPMI (Amaro & Bacao, 2024). Rentang nilai yang dihasilkan oleh coherence score adalah 0 sampai 1 dengan semakin besar nilai mengindikasikan topik lebih mudah dipahami. *UCI coherence* menghitung PMI antara semua pasangan kata dari satu topik menggunakan corpus eksternal (misalnya Wikipedia). Jika dua kata sering muncul bersama dalam corpus referensi, maka diasumsikan mereka punya makna yang berkaitan.

Cv Coherence merupakan salah satu metrik evaluasi topik yang paling banyak digunakan karena konsistensinya dengan penilaian manusia. *Cv Coherence* dikembangkan sebagai metrik gabungan yang mengintegrasikan beberapa pendekatan evaluasi sebelumnya, seperti penggunaan sliding window, co-occurrence kata, pembentukan vektor konteks, dan perhitungan cosine similarity antar pasangan kata dalam satu topik. *Cv Coherence* terbukti memberikan korelasi tertinggi dengan penilaian manusia, karena menggabungkan statistik *co-occurrence* lokal dan global, serta mempertimbangkan pembobotan semantik berbasis vektor (Röder dkk., 2015). Proses perhitungannya dimulai dengan membentuk pasangan kata (word pairs) dari kata-kata terpenting dalam satu topik, kemudian menghitung frekuensi kemunculan bersama antar kata tersebut dalam korpus referensi menggunakan teknik *sliding window*. Selanjutnya, untuk setiap kata dibentuk vektor konteks berdasarkan kata-kata tetangganya, dan antar pasangan kata dihitung nilai cosine similarity-nya. Skor Cv diperoleh sebagai rata-rata nilai cosine similarity dari semua pasangan kata tersebut, sehingga semakin tinggi skor Cv, semakin besar kesamaan semantik antar kata, dan semakin koheren topik yang dihasilkan. Nilai Cv berada dalam rentang antara 0 hingga 1. Skor di atas 0,50

dianggap menunjukkan topik yang sangat baik dan koheren secara semantik. Nilai di bawah 0,40 biasanya menunjukkan bahwa kata-kata dalam satu topik tidak cukup saling berkaitan dan topik tersebut sulit diinterpretasikan secara semantik. Penggunaan metrik ini tidak hanya terbatas untuk mengevaluasi kualitas topik, tetapi juga sangat berguna dalam menentukan jumlah topik optimal (K) dan membandingkan beberapa konfigurasi pemodelan topik seperti LDA, NMF, GSDMM maupun BERTopic (Amaro & Bacao, 2024). *Cv Coherence* dapat dihitung secara praktis menggunakan *library* seperti Gensim melalui objek CoherenceModel, dengan memasukkan model topik, kamus kata, dan korpus teks yang sudah diproses. *Cv Coherence* mengukur kesamaan semantik antar kata dalam sebuah topik menggunakan vektor konteks berbasis NPMI dan dihitung melalui cosine similarity antar vektor konteks kata seperti yang ditunjukkan pada persamaan 2.1 (Amaro & Bacao, 2024):

$$CV = \frac{1}{\binom{N}{2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \cos\_\text{sim}\left(v(w_i), v\left(w_j\right)\right) \quad (2.1)$$

dengan N adalah jumlah kata dalam satu topik yang akan dibandingkan (biasanya diambil top 10 kata dalam sebuah topik). Setiap kata w_i, w_j merupakan kata ke-i dan ke-j dalam topik diwakili oleh vector konteks $v(w_i), v\left(w_j\right)$ yang dihitung menggunakan NPMI (*Normalized Pointwise Mutual Information*), yang mengukur hubungan statistik antar kata dalam korpus kata tersebut terhadap kata lain dalam korpus/ dokumen. Cosine similarity atau cos_sim yaitu ukuran kesamaan antar dua vektor konteks, nilainya berada di antara -1 hingga 1, semakin dekat dengan 1 menunjukkan dua kata tersebut secara semantik memiliki kesamaan konteks yang tinggi.

Nilai Cv berada pada rentang 0 hingga 1. Semakin tinggi nilai CV, semakin baik koherensi topik tersebut. Dalam studi yang dilakukan oleh Widiantoro dkk (2023) yang menggunakan data ulasan pengguna pada aplikasi FinTech di Indonesia, nilai koherensi Cv digunakan sebagai dasar untuk menilai hasil pemodelan topik. Studi tersebut secara eksplisit menetapkan bahwa nilai ambang batas koherensi (threshold) ditetapkan sebesar 0.50, dengan rata-rata nilai koherensi topik yang dihasilkan mencapai 0.564. Topik-topik dengan nilai koherensi di atas

0.50 dianggap memiliki keterkaitan semantik yang cukup kuat (Widiantoro dkk., 2023).

Cv Coherence dapat dihitung secara praktis menggunakan library seperti Gensim melalui objek CoherenceModel, dengan memasukkan model topik, kamus kata, dan korpus teks yang sudah diproses. Validitas Cv sebagai metrik evaluasi topik terbaru yang menggunakan representasi vektor konteks untuk mengukur kesamaan semantik antara kata-kata dalam suatu topik, pendekatan tersebut menunjukkan bahwa Cv memiliki korelasi tertinggi dengan penilaian manusia dibanding metrik lainnya seperti UMass, UCI, dan NPMI sehingga sering digunakan sebagai acuan utama dalam penelitian terbaru (Amaro & Bacao, 2024; Röder dkk., 2015).

Topic coherence sebagai evaluasi semantik sangat berguna dalam pemodelan topik. Evaluasi ini tidak hanya membantu dalam menilai kualitas topik secara objektif tetapi juga memudahkan interpretasi hasil pemodelan topik oleh peneliti maupun praktisi di berbagai bidang. Metrik seperti Cv, UMass, dan UCI memberikan sudut pandang berbeda mengenai hubungan semantik dan kemunculan bersama antar kata, sehingga mampu menggambarkan sejauh mana topik yang dihasilkan mudah dipahami oleh manusia. Kekayaan evaluasi ini menjadi penting terutama pada analisis teks pendek seperti tweet, di mana informasi semantik sangat padat dan bergantung pada konteks. Dengan menggabungkan metrik koherensi tradisional dan pendekatan berbasis embedding, penelitian ini membangun fondasi evaluasi topik yang lebih menyeluruh, akurat, dan relevan terhadap karakteristik data media sosial

2.2.7 Perplexity

Perplexity adalah metrik evaluasi yang digunakan dalam pemodelan topik untuk mengukur seberapa baik model dapat memprediksi distribusi kata dalam suatu korpus teks (Lafferty dan Blei, 2005). Perplexity sering digunakan dalam model berbasis probabilistik, seperti *Latent Dirichlet Allocation* (LDA) dan *Gibbs Sampling Dirichlet Mixture Model* (GSDMM), untuk menilai kualitas model dalam merepresentasikan teks yang diolah. Semakin rendah nilai perplexity, semakin baik

model dalam memprediksi kata-kata dalam dokumen, yang berarti bahwa distribusi topik yang dihasilkan lebih sesuai dengan data yang ada.

Secara matematis, perplexity dihitung sebagai kebalikan dari kemungkinan rata-rata logaritmik sebuah kata dalam suatu dokumen, seperti yang ditunjukkan pada persamaan 2.2 (Blei dkk., 2002):

$$Perplexity(D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M} logP(w_d)}{\sum_{d=1}^{M} N_d}\right\} \quad (2.2)$$

dengan $P(w_d)$ adalah probabilitas suatu kata atau rangkaian kata dalam dokumen d berdasarkan model topik yang digunakan. *M* adalah jumlah dokumen dalam test set dan *N* adalah jumlah kata dalam korpus.

Dalam konteks pemodelan topik, perplexity sering digunakan untuk membandingkan beberapa model dengan parameter yang berbeda, misalnya jumlah topik atau metode representasi teks yang digunakan. Namun, perlu dicatat bahwa meskipun perplexity yang lebih rendah menunjukkan bahwa model lebih baik dalam menyesuaikan data, nilai ini tidak selalu berkorelasi dengan interpretabilitas topik yang dihasilkan. Model dengan perplexity rendah dapat menghasilkan topik yang sulit dipahami atau tidak memiliki makna yang jelas bagi manusia. Oleh karena itu, dalam penelitian ini, perplexity digunakan sebagai metrik komplementer bersama *coherence score*. Kombinasi kedua metrik ini memungkinkan evaluasi yang lebih seimbang antara akurasi probabilistik model dan interpretabilitas topik yang dihasilkan. Dengan menggunakan IndoBERTweet sebagai representasi teks dalam GSDMM, penelitian ini mengevaluasi apakah pendekatan ini dapat menurunkan perplexity sambil meningkatkan *coherence score*, sehingga menghasilkan analisis topik yang lebih akurat dan bermanfaat dalam memahami opini pelanggan di media sosial X.

2.2.8 Metrik Evaluasi Semantik

Evaluasi kualitas topik merupakan aspek penting dalam pemodelan topik untuk memastikan bahwa kumpulan kata yang muncul pada setiap topik tidak hanya koheren secara statistik, tetapi juga bermakna secara semantik. Selama ini, kualitas topik umumnya dinilai menggunakan metrik koherensi tradisional seperti Cv, UMass, dan UCI. Meskipun metrik-metrik tersebut telah digunakan secara luas,

terdapat sejumlah keterbatasan yang menyebabkan penilaiannya tidak selalu mencerminkan kedekatan makna yang sebenarnya antar kata dalam topik. Oleh karena itu, pendekatan berbasis word embedding seperti Word2Vec, BERT, dan IndoBERTweet semakin banyak digunakan untuk memberikan evaluasi semantik yang lebih akurat. Pendekatan ini memungkinkan representasi makna kata dalam bentuk vektor dan pengukuran keserupaan semantik menggunakan cosine similarity, sehingga dapat memberikan gambaran yang lebih mendalam mengenai kualitas makna topik.

Untuk mengatasi keterbatasan tersebut, pendekatan evaluasi semantik berbasis word *embedding* diperkenalkan untuk menyediakan penilaian yang lebih kaya secara linguistik. Dalam evaluasi semantik, kedekatan makna antar kata dalam satu topik diukur menggunakan cosine similarity, yaitu ukuran kesamaan arah dua vektor dalam ruang *embedding* (Jurafsky dan Martin, 2025). Cosine similarity bernilai 1 menunjukkan bahwa dua kata sangat mirip maknanya, sedangkan nilai mendekati 0 menunjukkan tidak adanya hubungan semantik, seperti yang ditunjukkan pada persamaan 2.3. (Yadav dkk., 2023)

$$\textit{Cosine Similarity}(u, v) = \frac{u \cdot v}{\|u\|\|v\|} \tag{2.3}$$

dengan $u.v$ adalah hasil perkalian titik dua representasi vector, $\|u\|$ dan $\|v\|$ adalah panjang masing masing vektor. Dengan teknik ini, kualitas topik dinilai melalui beberapa indikator WCF, BCF dan CF (Abdelmotaleb dkk., 2023) :

1. *Within-Cluster Variance* (WCV) untuk mengukur seberapa kohesif/ dekat dokumen dalam satu topik;

$$WCV = \frac{\sum \textit{Cosine Distance Antar-Pasangan}}{\textit{Jumlah Pasangan}} \tag{2.4}$$

   dengan menjumlahkan seluruh *cosine distance* dari setiap pasangan dokumen terdapat di dalam klaster, kemudian hasil akumulasi tersebut dinormalisasi dengan membaginya terhadap jumlah total pasangan yang terbentuk. Hasil akhir dari persamaan ini merepresentasikan rata-rata jarak ketidaksamaan antar elemen; di mana nilai WCV yang semakin rendah mengindikasikan bahwa dokumen-dokumen di dalam klaster tersebut memiliki tingkat kemiripan yang tinggi.

2. *Between-Cluster Variance* (BCV) untuk melihat seberapa jauh antar topik berbeda secara semantik; dan

$$BCV = \frac{\sum Cosine\ Distance\ Antar\ Centroid}{Jumlah\ Pasangan\ Centroid} \quad (2.5)$$

dengan menjumlahkan *cosine distance* antara titik pusat (*centroid*) dari setiap pasangan klaster, kemudian hasil akumulasi tersebut dinormalisasi dengan membaginya terhadap jumlah total pasangan centroid yang terbentuk. Hasil akhir dari persamaan ini merepresentasikan rata-rata jarak perbedaan antar topik; di mana nilai BCV yang semakin tinggi mengindikasikan bahwa klaster-klaster yang terbentuk memiliki perbedaan karakteristik yang jelas dan terpisah dengan baik satu sama lain.

3. *Cost Function* (CF) sebagai kombinasi keduanya untuk memberikan ukuran stabil yang mencerminkan keseimbangan kedekatan dan separasi topik

$$CF = \lambda_{cf} \cdot WCV - (1 - \lambda_{cf}) \cdot BCV \quad (2.6)$$

dengan mengintegrasikan dua metrik evaluasi, yaitu *Within-Cluster Variation* (WCV) dan *Between-Cluster Variation* (BCV), melalui mekanisme pembobotan linier. Variabel lambda bertindak sebagai *hyperparameter* pengatur yang menyeimbangkan prioritas antara kekompakan internal klaster (WCV) dan keterpisahan antar klaster (BCV).

Berbeda dari koherensi tradisional, metrik-metrik ini menilai kualitas topik berdasarkan makna aktual dalam *embedding*, bukan sekadar kemunculan kata. Secara keseluruhan, metrik evaluasi semantik menawarkan perspektif yang lebih mendalam terhadap kualitas topik, terutama pada dataset teks pendek seperti tweet. Dengan menggabungkan kekuatan *embedding* IndoBERTweet dan cosine similarity, penilaian topik menjadi lebih mendekati persepsi manusia dan lebih relevan dengan konteks bahasa Indonesia di media sosial. Dengan demikian, integrasi metrik koherensi tradisional dan evaluasi berbasis *embedding* memberikan gambaran yang lebih luas terhadap kualitas model topik.