BAB I
PENDAHULUAN

1.1 Latar Belakang

Tidak bisa dipungkiri bahwa media sosial telah menjadi bagian tak terpisahkan dalam kehidupan generasi milenial dan generasi Z saat ini. Hampir semua individu dari kedua generasi tersebut memiliki kebutuhan yang khusus terhadap media sosial, baik untuk tujuan mencari informasi, mengeksplorasi pengalaman baru, atau bahkan sebagai sumber hiburan untuk mengurangi stres. Selain itu, media sosial juga berfungsi sebagai wadah untuk menyuarakan aspirasi, karena di dalamnya terdapat audiens yang siap mendengarkan dan merespons [1]. Salah satu media sosial yang cukup populer adalah X. Di Indonesia sendiri terdapat kurang lebih 24.69 juta pengguna aktif terhitung hingga pertengahan tahun 2024 ini [2]. X ini seringkali digunakan sebagai sarana bertukar opini dan informasi serta sarana penyampaian pendapat dan aspirasi.

Pendapat dan opini orang lain akan sangat berguna untuk mendapatkan *insight* dari perspektif yang beragam. Tidak terkecuali bagi Tim Nasional Sepakbola Indonesia usia 23 yang telah berlaga di Piala Asia U-23 2024. Staf kepelatihan Timnas, bisa saja menggali informasi dari X untuk menjadi bahan evaluasi. Banyak penggemar dan pendukung Timnas yang memberikan pendapat dan sarannya melalui postingan di X. Postingan-postingan tersebut harus dianalisis sehingga dapat diketahui topik-topik apa saja yang ramai dibahas oleh para penggemar. Oleh karena itu dipilihlah kata kunci #TimnasDay pada penelitian ini, dengan harapan topik-topik yang dihasilkan dari proses pemodelan topik ini dapat menjadi *insight* atau informasi yang berguna.

Postingan-postingan pada X, dikumpulkan dan dijadikan dataset untuk penelitian ini karena platform ini menawarkan karakteristik unik yang sangat relevan dengan tujuan penelitian, yaitu menganalisis teks pendek yang sarat dengan bahasa tidak baku, slang, dan elemen informal khas media sosial. X, sebagai salah satu platform dengan basis pengguna yang luas dan dinamis, secara aktif menampilkan diskusi dan opini seputar peristiwa penting, termasuk dukungan dan respon terhadap #TimnasDay, yang mencerminkan semangat dan masukan-

masukan dari masyarakat. Dengan memilih X sebagai sumber data, penelitian ini dapat menangkap secara autentik dinamika komunikasi dan perilaku pengguna dalam konteks nyata, yang mungkin tidak akan tersaji jika data diambil dari media sosial lain atau sumber lain dengan karakteristik berbeda.

Dengan postingan yang sangat banyak, tidak mungkin untuk melakukan analisis secara manual dan satu persatu. Oleh karena itu diperlukan teknik pemodelan topik (*topic modelling*) untuk menganalisis kumpulan postingan yang sangat banyak (korpus). Pemodelan topik adalah pendekatan yang sering digunakan untuk menemukan pola semantik tersembunyi dalam sebuah korpus teks dan secara otomatis mengidentifikasi topik-topik yang ada di dalamnya. Teknik ini merupakan jenis pemodelan statistik yang memanfaatkan *unsupervised machine learning* untuk menganalisis dan mengidentifikasi kelompok-kelompok kata serupa dalam suatu teks [3].

Pemodelan topik dipilih karena metode ini mampu mengungkap struktur laten dan pola semantik dalam data teks yang kompleks, sehingga memberikan wawasan mendalam yang tidak bisa diperoleh hanya dengan penghitungan kata biasa. Sementara penghitungan frekuensi kata hanya mengukur seberapa sering kata muncul, pemodelan topik dapat mengidentifikasi hubungan antar kata dan mengelompokkan dokumen berdasarkan tema atau topik tersembunyi, yang sangat penting dalam konteks data pendek dan penuh dengan bahasa tidak baku serta slang seperti pada postingan di X. Dengan demikian, pemodelan topik memberikan gambaran yang lebih komprehensif dan bermakna mengenai konten serta dinamika komunikasi pengguna, yang pada akhirnya mendukung pengambilan keputusan berbasis data dengan lebih efektif.

Ada beberapa algoritma yang umum digunakan dalam pemodelan topik, salah satunya adalah *Latent Dirichlet Allocation* (LDA). LDA mengasumsikan bahwa setiap dokumen merupakan kombinasi dari beberapa topik yang mendasar, dan setiap topik dianggap sebagai campuran dari distribusi topik yang tak terbatas. Hasilnya bersifat tidak pasti karena dapat berbeda setiap kali model dijalankan, bahkan dengan dataset yang sama. Sementara algoritma BERTopic merupakan teknik yang cukup baru untuk pemodelan topik. Teknik ini menggunakan *Bidirectional Encoder Representations from Transformers* (BERT) *embeddings*

dari Google dan variasi berbasis kelas dari *Term Frequency - Inverse Document Frequency* (TF-IDF) [4].

Pemodelan topik pada dasarnya termasuk dalam kategori *clustering* karena tujuannya adalah mengelompokkan dokumen atau kata-kata berdasarkan karakteristiknya. Namun, algoritma yang digunakan dalam pemodelan topik, seperti LDA atau BERTopic, berbeda dengan teknik *clustering* tradisional (misalnya, *K-Means*). LDA, misalnya, melakukan *soft clustering* dengan memberikan distribusi probabilitas dari beberapa topik untuk setiap dokumen, sedangkan BERTopic menggabungkan representasi semantik (dari model seperti BERT) dengan algoritma *clustering* (misalnya, HDBSCAN) untuk mengelompokkan dokumen. Dengan demikian, meskipun keduanya memiliki esensi mengelompokkan data, cara kerja dan output yang dihasilkan pun berbeda.

Latent Dirichlet Allocation (LDA) banyak digunakan dalam pemodelan topik karena kemampuannya yang efektif dalam mengidentifikasi struktur tematik tersembunyi dalam kumpulan data teks yang besar dan tidak terstruktur. Sebagai model generatif probabilistik, LDA mengasumsikan bahwa setiap dokumen terdiri dari campuran beberapa topik, dengan setiap topik diwakili oleh distribusi kata tertentu. Hal ini memungkinkan LDA untuk menangani kompleksitas dan variasi dalam bahasa alami secara efisien. Fleksibilitas ini menjadikan LDA sangat cocok untuk berbagai aplikasi, mulai dari penelitian akademik hingga analisis media sosial, karena dapat mengelompokkan konten secara otomatis tanpa memerlukan data berlabel [5]. Selain itu, sifat *unsupervised* pada LDA memungkinkan algoritma ini untuk mengidentifikasi topik tanpa memerlukan pengetahuan awal mengenai subjek yang dianalisis, menjadikannya alat yang sangat efektif untuk analisis data eksploratif [6]. Secara keseluruhan, perpaduan antara adaptabilitas, efisiensi, dan kemudahan implementasi LDA menjadikannya sebagai salah satu pilihan utama dalam pemodelan topik [7].

Sementara BERTopic semakin populer untuk pemodelan topik berkat pendekatan inovatifnya yang menggabungkan keunggulan model transformer dengan teknik pengelompokan (*clustering*), menghasilkan topik yang mudah diinterpretasikan dan bermakna. Metode ini menggunakan *embedding* BERT bersama dengan *class-based* TF-IDF (c-TF-IDF) untuk membentuk klaster

dokumen yang padat, yang memungkinkan ekstraksi topik dengan kata kunci yang signifikan sambil tetap menjaga kejelasan deskripsi topik. Pendekatan ini tidak hanya meningkatkan interpretabilitas topik, tetapi juga mendukung analisis tema tersembunyi pada berbagai kumpulan data, menjadikannya alat yang berguna dalam berbagai bidang seperti analisis media sosial, survei pelanggan, dan penelitian akademik. Kemampuan BERTopic untuk memvisualisasikan hirarki topik, menjadikan BERTopic alat yang kuat untuk mengungkap wawasan dari data teks yang tidak terstruktur [8].

Membandingkan antara *Latent Dirichlet Allocation* (LDA) dan BERTopic menjadi penting untuk memahami perkembangan teknik pemodelan topik, terutama karena keduanya memenuhi kebutuhan analisis dan karakteristik data yang berbeda. LDA, sebagai model probabilistik tradisional, unggul dalam menangani kumpulan teks berukuran besar dan menawarkan landasan yang kokoh dalam mengidentifikasi topik laten melalui pendekatan generatifnya. Namun, LDA memiliki beberapa keterbatasan, seperti perlunya menetapkan jumlah topik di awal dan kurangnya sensitivitas terhadap urutan kata serta konteks *semantic*[9]. Di sisi lain, BERTopic menggunakan *embedding* berbasis *transformer* untuk menangkap konteks dan hubungan semantik dalam teks, memungkinkan representasi topik yang lebih koheren dan mudah diinterpretasi. Kemampuannya dalam membentuk klaster yang padat serta menghasilkan topik secara adaptif berdasarkan data menjadikannya sangat cocok untuk aplikasi modern, seperti analisis media sosial dan konten dinamis [10]. Penelitian ini berfokus pada membandingkan algoritma LDA dan BERTopic.

Perbandingan kedua algoritma ini sudah dilakukan pada penelitian sebelumnya. Namun, penelitian ini mengambil pendekatan berbeda dengan menggunakan kumpulan tweet berbahasa Indonesia, yang memiliki karakteristik unik berupa dokumen pendek, struktur kalimat yang tidak baku, dan penggunaan kata-kata informal. Karakteristik ini menimbulkan tantangan tambahan dalam pemodelan topik dibandingkan dengan korpus artikel berita yang lebih terstruktur. Selain itu, penelitian ini mengkaji performa LDA dengan tiga variasi parameter passes (20, 50, dan 100) untuk mengeksplorasi pengaruh jumlah iterasi terhadap hasil pemodelan.

Hasil dari kedua algoritma tersebut akan dievaluasi menggunakan metrik *Topic Coherence* dan *Topic Diversity*. *Topic Coherence* akan mengevaluasi kualitas topik yang dihasilkan dengan mengukur keterkaitan antar kata pada topik. Sedangkan *Topic Diversity* akan mengevaluasi keunikan antara topik yang dihasilkan oleh kedua algoritma ini nantinya. Metrik ini akan menilai seberapa berbeda antar topik yang dihasilkan satu sama lain. Terakhir topik-topik yang nantinya dihasilkan oleh kedua algoritma ini, akan dievaluasi kualitasnya secara manual.

Urgensi penelitian ini terletak pada kebutuhan untuk menentukan algoritma yang paling efektif dalam pemodelan topik pada data dengan karakteristik unik, seperti postingan di platform X, yang umumnya berisi teks pendek, kata-kata tidak baku, slang, hingga kata-kata yang disingkat. Kondisi ini seringkali menjadi tantangan dalam analisis teks, sehingga memilih algoritma yang tepat sangat penting untuk menghasilkan topik yang relevan dan bermakna. Dengan membandingkan kinerja *Latent Dirichlet Allocation* (LDA) dan BERTopic, penelitian ini diharapkan memberikan panduan bagi peneliti atau praktisi yang ingin melakukan pemodelan topik pada data serupa, sehingga dapat memaksimalkan hasil analisis dan pengambilan keputusan berbasis teks secara lebih efektif.

1.2 Perumusan Masalah

Berdasarkan hasil pemaparan latar belakang diatas, didapat rumusan masalah sebagai berikut:

1. Bagaimana menerapkan algoritma LDA dan BERTopic pada pemodelan topik?
2. Berapa nilai *Topic Coherence* dan *Topic Diversity* dari algoritma LDA dan BERTopic?
3. Bagaimana kualitas topik yang dihasilkan dari algoritma LDA dan BERTopic?

1.3 Batasan Masalah

Batasan masalah akan membuat penyusunan dan pembahasan penelitian ini lebih terarah. Berikut merupakan Batasan masalah yang digunakan pada penelitian ini:

1. Algoritma yang digunakan pada penelitian ini adalah LDA dan BERTopic.
2. Korpus yang akan digunakan adalah postingan berbahasa Indonesia pada X dengan kata kunci #TimnasDay dan diposting dalam jangka waktu selama pelaksanaan Piala Asia U-23 AFC 2024.
3. Data berjumlah 1.000 cuitan di aplikasi X.
4. Penelitian ini menggunakan bahasa pemrograman Python.
5. Hasil penelitian menampilkan hasil pemodelan topik dari kedua algoritma yaitu LDA dan BERTopic.

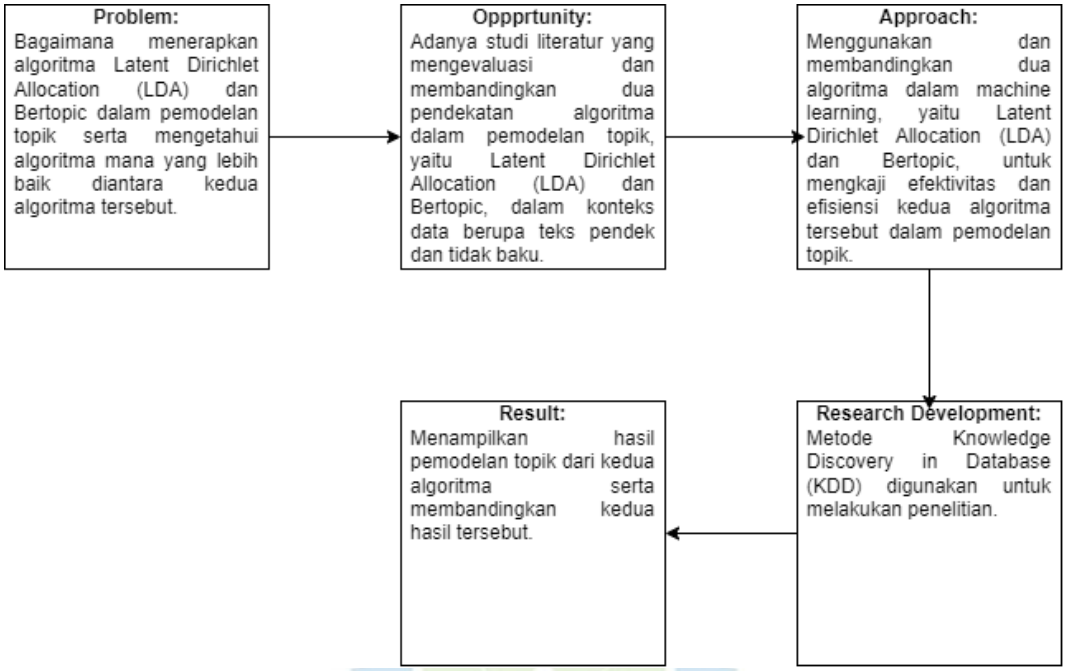
1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang dipaparkan didapat tujuan yang ingin dicapai dalam penelitian ini sebagai berikut:

1. Menerapkan algoritma LDA dan BERTopic pada pemodelan topik.
2. Mengetahui nilai *Topic Coherence* dan *Topic Diversity* dari algoritma LDA dan BERTopic.
3. Mengetahui topik yang dihasilkan dari algoritma LDA dan BERTopic.

1.5 Kerangka Pemikiran Penelitian

Kerangka pemikiran penelitian ini diilustrasikan dalam Gambar 1.1. Studi ini membahas kompleksitas yang timbul dari beragamnya algoritma pemodelan topik yang tersedia. Untuk mengatasi permasalahan ini, penelitian mengusulkan pendekatan komparatif antar berbagai algoritma. Lebih lanjut, penelitian ini mengaplikasikan dua algoritma yang berbeda dalam konteks pemodelan topik. Proses ini memanfaatkan metodologi *Knowledge Discovery in Database* (KDD) untuk pengembangan sistemnya. Setiap langkah dalam metode KDD diimplementasikan menggunakan bahasa pemrograman Python. Tujuan utama dari penelitian ini adalah untuk mengidentifikasi algoritma yang memiliki performa terbaik dalam pemodelan topik, berdasarkan hasil analisis komparatif yang dilakukan.



Gambar 1. 1 Kerangka Pemikiran

1.6 Metodologi Penelitian

Metodologi penelitian ini mengikuti alur *Knowledge Discovery in Database* (KDD), yang terdiri dari beberapa tahapan utama. Penelitian diawali dengan menentukan *knowledge discovery goal*, di mana tujuan utamanya adalah untuk membandingkan dua algoritma pemodelan topik, yaitu *Latent Dirichlet Allocation* (LDA) dan BERTopic, dalam menganalisis topik-topik yang muncul dari cuitan terkait Timnas Indonesia selama pergelaran AFC U23 2024. Data yang digunakan adalah kumpulan cuitan berbahasa Indonesia yang diperoleh dari aplikasi X menggunakan kata kunci "#TimnasDay." Penentuan tujuan ini menjadi dasar untuk langkah-langkah selanjutnya dalam proses KDD.

Tahap kedua adalah integrasi data. Pada tahap ini, data cuitan dikumpulkan dari X dengan memanfaatkan *tweet-harvest* untuk melakukan pencarian berdasarkan kata kunci yang relevan, yaitu "#TimnasDay," dalam jangka waktu yang sesuai dengan periode pergelaran AFC U23 2024. Data yang dikumpulkan meliputi teks cuitan, informasi waktu dan pengguna, dan lain-lain. Data mentah ini kemudian diintegrasikan dan disimpan dalam format yang siap untuk dianalisis.

Setelah data berhasil diintegrasikan, langkah pra-pemrosesan dilakukan untuk membersihkan dan menyiapkan data agar sesuai dengan kebutuhan

pemodelan. Teks cuitan diproses melalui beberapa langkah seperti penghapusan tanda baca, angka, *stopwords*, normalisasi kata dan lain-lain agar dapat dipahami oleh algoritma. Selanjutnya, dilakukan tokenisasi dan *lemmatization* untuk mengurangi kompleksitas teks. Setelah data dipersiapkan, pemodelan dilakukan dengan dua algoritma, yaitu LDA dan BERTopic, untuk memodelkan topik dari cuitan. Hasil pemodelan dari kedua algoritma tersebut kemudian dibandingkan berdasarkan akurasi serta kualitas dan keberagaman topik yang dihasilkan.

1.7 Sistematika Penulisan

Pada penelitian ini, penulis membagi sistematika penulisan menjadi 5 bab sebagai berikut:

BAB I: PENDAHULUAN

Dalam BAB I, penulis memaparkan mengenai latar belakang penelitian, rumusan masalah, tujuan penelitian, batasan masalah, kerangka berpikir, dan memaparkan bagaimana sistematika penulisannya.

BAB II: KAJIAN LITERATUR

Dalam BAB II, penulis memaparkan mengenai landasan teori serta penelitian terdahulu yang menjadi acuan dalam penelitian ini.

BAB III: METODE PENELITIAN

Dalam BAB III, penulis memaparkan metode yang akan digunakan dalam penelitian ini, meliputi metode penelitian analisi sumber data dan metode pengembangan sistem.

BAB IV: HASIL DAN PEMBAHASAN

Dalam BAB IV, penulis memaparkan hasil penelitian yang sudah dijalankan serta menjelaskan secara rinci mengenai evaluasi model yang sudah didapatkan.

BAB V: KESIMPULAN DAN SARAN

Dalam BAB V, penulis memaparkan mengenai kesimpulan yang sudah didapatkan serta memberi saran untuk penelitian selanjutnya.