# LAPORAN

# PENERAPAN PEMODELAN TOPIK PADA MEDIA SOSIAL TWITTER MENGGUNAKAN ALGORITMA LATENT DIRICHLECT ALLOCATION

## RISET INFORMATIKA D081

**Dosen pengampu :**

Dr. Basuki Rahmat, S.Si., M.T.

**Dibuat Oleh:**

21081010286 Azila Lailannafisa

**PROGRAM STUDI INFORMATIKA**

**FAKULTAS ILMU KOMPUTER**

**UPN "VETERAN" JAWA TIMUR**

**2024**

# BAB 1

# PENDAHULUAN

## 1.1 Latar Belakang

Media sosial telah berkembang menjadi salah satu alat utama bagi masyarakat untuk berbagi informasi, menyampaikan aspirasi, dan memberikan opini mengenai berbagai isu sosial dan kebijakan pemerintah. Platform seperti Twitter memungkinkan percakapan yang cepat dan terbuka, menjadikannya sumber data yang kaya untuk memahami persepsi masyarakat. Sebagai salah satu negara dengan pengguna media sosial terbesar, 57,5% dari populasi di Indonesia aktif menggunakan Twitter, menjadikannya platform penting untuk mengamati tren opini publik dan umpan balik terhadap kebijakan pemerintah (Arianto, 2021)

Dalam konteks pengelolaan kota, Pemerintah Kota Surabaya, sebagai salah satu kota metropolitan terbesar di Indonesia, menghadapi tantangan besar dalam memahami opini masyarakat yang seringkali terekspresi melalui media sosial. Meskipun platform ini menyediakan data dalam jumlah besar, formatnya yang tidak terstruktur membuat analisis manual tidak efisien. Oleh karena itu, diperlukan teknik analisis big data, seperti pemodelan topik, untuk membantu mengolah data ini secara otomatis dan memberikan wawasan yang relevan.

Salah satu metode pemodelan topik yang banyak digunakan adalah Latent Dirichlet Allocation (LDA). LDA adalah algoritma berbasis probabilitas yang mampu mengidentifikasi "topik" tersembunyi dalam kumpulan data teks besar. Penelitian terdahulu telah menunjukkan bahwa LDA efektif untuk mengeksplorasi data teks besar, seperti komentar media sosial, untuk menghasilkan topik yang relevan dan dapat diinterpretasikan dengan baik (Blei, 2003). Metode ini dapat membantu pemerintah memahami isu-isu yang sedang dibicarakan masyarakat dan memberikan wawasan yang berguna untuk mendukung pengambilan keputusan kebijakan (Alghamdi & Alfalqi, 2015).

Penelitian di berbagai negara telah menunjukkan keberhasilan pemodelan topik menggunakan LDA untuk analisis media sosial. Contohnya, studi pada data Twitter selama bencana alam di Amerika Serikat menunjukkan kemampuan LDA dalam mengidentifikasi topik yang mencerminkan kebutuhan masyarakat dalam situasi krisis (Resch, 2018). Selain itu, penelitian lain menunjukkan bahwa analisis berbasis LDA efektif dalam mengidentifikasi tema utama dalam umpan balik masyarakat terhadap layanan pemerintah di Inggris, yang kemudian membantu dalam menyusun kebijakan yang lebih responsif (Xiao, 2016).

Dalam konteks Indonesia, penelitian sebelumnya yang dilakukan di Pemerintah Kota Malang menggunakan LDA telah berhasil mengidentifikasi isu-isu utama seperti banjir, hukum, wisata, dan proyek pemerintah dari komentar masyarakat

di Twitter (Zainiyah, 2017). Namun, penelitian ini menunjukkan bahwa penggunaan langkah preprocessing seperti stemming dapat memengaruhi kualitas hasil topik, dengan model tanpa stemming menghasilkan topik yang lebih koheren (Rahmawati, 2021).

Oleh karena itu, penelitian ini bertujuan untuk menerapkan metode LDA dalam analisis komentar masyarakat di Twitter terkait Pemerintah Kota Surabaya. Penelitian ini diharapkan dapat menghasilkan wawasan berharga mengenai isu-isu utama yang menjadi perhatian masyarakat, yang dapat digunakan untuk mendukung pengambilan keputusan berbasis data dan meningkatkan pelayanan publik.

## 1.2 Rumusan Masalah

Permasalahan yang dapat dirumuskan dari latar belakang tersebut antara lain:

1) Bagaimana pola topik komentar masyarakat di media sosial Twitter terkait Pemerintahan Kota Surabaya dapat diidentifikasi menggunakan metode Latent Dirichlet Allocation (LDA)?
2) Apa saja topik utama yang muncul dari analisis komentar masyarakat di Twitter terhadap kebijakan dan aktivitas Pemerintahan Kota Surabaya?
3) Bagaimana tingkat relevansi dan representasi topik yang dihasilkan oleh model LDA dalam memahami sentimen serta isu yang sering dibahas oleh masyarakat?
4) Bagaimana hasil pemodelan topik dapat dimanfaatkan oleh Pemerintahan Kota Surabaya untuk meningkatkan pelayanan publik dan respons terhadap aspirasi masyarakat?

## 1.3 Tujuan Penelitian

Berdasarkan rumusan masalah tersebut, tujuan yang ini dicapai pada penelitian ini adalah:

1) Mengidentifikasi topik utama dalam komentar masyarakat di media sosial Twitter yang terkait dengan Pemerintah Kota Surabaya menggunakan metode Latent Dirichlet Allocation (LDA).
2) Mengevaluasi pengaruh proses stemming (mengembalikan kata ke bentuk dasar) terhadap performa model LDA, termasuk nilai perplexity dan coherence score.
3) Menganalisis efektivitas parameter model LDA (seperti jumlah topik, nilai alpha, dan beta) dalam menghasilkan topik yang relevan, koheren, dan mudah diinterpretasikan.
4) Memberikan wawasan tentang persepsi masyarakat terhadap isu-isu sosial dan kebijakan Pemerintah Kota Surabaya, yang dapat digunakan untuk mendukung pengambilan keputusan berbasis data.

## 1.4 Manfaat Penelitian

Dalam pelaksanaannya penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1) Memberikan wawasan berbasis data terkait aspirasi masyarakat untuk meningkatkan pelayanan public

2) Membantu Pemerintah Kota Surabaya memahami isu-isu utama yang dibicarakan masyarakat di Twitter untuk mendukung pengambilan keputusan.
3) Menjadi referensi metode pemodelan topik (LDA) dalam analisis data media sosial.
4) Meningkatkan efisiensi dalam mengolah data besar secara otomatis untuk menghasilkan informasi yang relevan.

## 1.5 Batasan Penelitian

1. Penelitian ini hanya menggunakan data komentar masyarakat yang diperoleh dari media sosial Twitter, yang secara spesifik menyebut atau terkait dengan Pemerintah Kota Surabaya.
2. Penelitian terbatas pada penggunaan metode Latent Dirichlet Allocation (LDA) untuk pemodelan topik, tanpa membandingkan dengan metode pemodelan topik lainnya.
3. Penelitian hanya mencakup proses preprocessing teks seperti stemming, case folding, tokenizing, dan filtering untuk meningkatkan kualitas data teks. Bahasa yang dianalisis dibatasi pada Bahasa Indonesia, tanpa mempertimbangkan komentar dalam bahasa lain.
4. Data yang digunakan dibatasi pada rentang waktu tertentu sesuai dengan periode pengambilan data melalui proses crawling di Twitter, tanpa mempertimbangkan data historis atau data masa depan.

# BAB 2

# TINJAUAN PUSTAKA

## 2.1 Penelitian Terdahulu

Banyak penelitian telah menggunakan LDA untuk analisis teks di berbagai domain. Misalnya, studi tentang analisis kebijakan publik sering menggunakan LDA untuk mengidentifikasi topik yang muncul dalam diskusi masyarakat. Selain itu, dalam bidang pemasaran, LDA digunakan untuk memahami kebutuhan dan preferensi pelanggan berdasarkan ulasan produk di media sosial.

Penelitian sebelumnya menunjukkan bahwa LDA efektif dalam mengidentifikasi pola-pola umum dalam kumpulan dokumen yang besar. Namun, penerapan LDA pada teks pendek, seperti tweet, sering kali menghadapi tantangan dalam hal kualitas hasil yang dihasilkan. Beberapa pendekatan, seperti pembentukan pseudo-dokumen, telah diusulkan untuk mengatasi masalah ini.

Penelitian oleh Budianto et al. (2023) menggunakan LDA untuk menganalisis tweet terkait isu kesehatan di Surabaya, terutama selama pandemi COVID-19. Penelitian ini bertujuan untuk mengidentifikasi topik-topik kesehatan yang menjadi perhatian masyarakat dan bagaimana informasi kesehatan disebarkan melalui Twitter. Hasil penelitian menunjukkan bahwa LDA efektif dalam mengelompokkan tweet berdasarkan tema kesehatan, memberikan wawasan tentang bagaimana masyarakat merespons isu kesehatan yang sedang berlangsung.

Rahman et al. (2021) menggunakan LDA untuk mengidentifikasi tren topik di Twitter selama pandemi COVID-19. Penelitian ini berfokus pada bagaimana masyarakat merespons situasi krisis melalui media sosial, dengan menganalisis tweet yang berkaitan dengan kesehatan, ekonomi, dan kebijakan pemerintah. Hasilnya menunjukkan bahwa LDA dapat mengungkapkan perubahan dalam perhatian publik terhadap berbagai isu seiring berjalannya waktu, serta bagaimana topik-topik tertentu mendominasi diskusi di Twitter.

Penelitian mengenai pemodelan topik di Twitter telah dilakukan dengan menggunakan berbagai algoritma selain Latent Dirichlet Allocation (LDA). Zhang et al. (2023) menerapkan Non-Negative Matrix Factorization (NMF) untuk menganalisis tweet terkait perubahan iklim, menunjukkan bahwa NMF dapat mengidentifikasi topik-topik yang lebih halus, seperti dampak lokal dari isu tersebut.

Sementara itu, Kumar dan Singh (2022) menggunakan Latent Semantic Analysis (LSA) untuk menganalisis tweet tentang kesehatan mental, di mana LSA membantu

dalam mengidentifikasi hubungan antara kata-kata dan konteks, sehingga dapat mengelompokkan tweet berdasarkan tema yang lebih kompleks.

## 2.2 Penerapan LDA dalam Media Sosial

Penerapan LDA pada media sosial telah menghasilkan berbagai temuan menarik. Misalnya, penelitian yang menggunakan LDA untuk menganalisis tweet terkait bencana alam berhasil mengidentifikasi tema-tema seperti bantuan, evakuasi, dan kerusakan. Penelitian lain menggunakan LDA untuk menganalisis tren mode dan gaya hidup berdasarkan tweet yang mengandung hashtag tertentu.

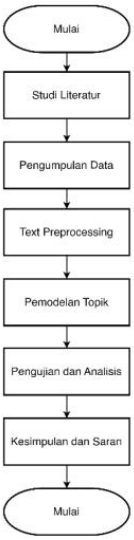
Namun, aplikasi LDA pada media sosial memerlukan perhatian khusus terhadap preprocessing data, seperti penghapusan stopwords, stemming, dan pengelompokan teks pendek. Pendekatan ini memastikan bahwa data yang dianalisis lebih representatif terhadap pola yang sebenarnya.

## 2.3 Kota Surabaya

Penelitian terkait analisis media sosial di Surabaya masih relatif terbatas. Beberapa studi telah mengeksplorasi penggunaan media sosial untuk memahami sentimen publik terhadap kebijakan lokal atau mengidentifikasi isu-isu yang sering dibicarakan, seperti kemacetan dan kebersihan kota. Namun, belum banyak penelitian yang secara khusus menggunakan LDA untuk mengidentifikasi tema-tema utama dari percakapan Twitter tentang Surabaya.

# BAB 3

# METODOLOGI




## 3.1 Pengumpulan data Menggunakan Metode Crawling

Pengumpulan data dilakukan dengan metode *crawling* untuk mengunduh komentar atau cuitan dari Twitter. Untuk keperluan ini, digunakan alat atau pustaka pemrograman seperti Python dengan bantuan *Twitter API* melalui pustaka seperti tweepy atau *scraping tools* yang relevan. Kata kunci atau tagar tertentu yang relevan dengan tema penelitian (misalnya "#politik" atau "#ekonomi") digunakan sebagai parameter pencarian data. Proses *crawling* diatur agar data yang diambil mencakup teks cuitan, *timestamp*, dan metadata seperti jumlah *likes* atau *retweets*.

Peneliti menentukan rentang waktu pengumpulan data (contohnya, selama 6 bulan terakhir) untuk memastikan data yang terkumpul cukup representatif. Selama proses *crawling*, peneliti mematuhi kebijakan Twitter terkait pengumpulan data agar proses berjalan secara etis dan legal. Setelah data berhasil dikumpulkan, dilakukan penyaringan awal untuk menghapus komentar yang redundan seperti *retweets*, spam, atau cuitan yang tidak relevan. Jika volume data terlalu besar, dilakukan pengambilan sampel acak untuk memastikan data yang dianalisis tetap representatif namun mudah dikelola.

## 3.2 Preprocessing

Setelah data berhasil dikumpulkan melalui proses *crawling*, langkah berikutnya adalah *preprocessing* data untuk memastikan data dalam format yang siap dianalisis menggunakan model LDA. Langkah pertama adalah membersihkan data dengan menghapus elemen-elemen yang tidak relevan seperti *stopwords* (kata-kata umum seperti "dan", "di", "ke"), URL, tagar, nama

pengguna, angka, serta simbol atau karakter khusus lainnya. Proses ini memastikan bahwa hanya informasi penting yang tersisa dalam data.

Data kemudian diproses lebih lanjut melalui tokenisasi, yaitu memecah teks cuitan menjadi unit-unit kata kecil (*tokens*). Selanjutnya, dilakukan stemming untuk mengubah kata-kata menjadi bentuk dasarnya menggunakan algoritma seperti Sastrawi untuk data berbahasa Indonesia. Setelah itu, data teks diubah menjadi representasi numerik menggunakan metode seperti *Bag-of-Words* atau *TF-IDF* agar dapat dianalisis oleh algoritma LDA.

### 3.3 Penerapan Pemodelan Topik

Pemodelan topik dilakukan menggunakan algoritma *Latent Dirichlet Allocation* (LDA). Implementasi dilakukan dengan menggunakan Python dan pustaka seperti gensim dan scikit-learn. Model LDA dirancang untuk mendeteksi topik-topik utama dari kumpulan komentar yang telah diproses.

Pada tahap awal, peneliti menentukan jumlah topik (contoh: 5, 10, atau 15 topik) berdasarkan literatur atau eksperimen awal. Model LDA kemudian dilatih untuk menghasilkan distribusi topik, yaitu daftar kata-kata dominan yang membentuk setiap topik. Proses ini juga mencakup pengaturan parameter seperti jumlah iterasi dan nilai *alpha* untuk memastikan hasil yang optimal.

### 3.4 Evaluasi Model

Evaluasi model dilakukan untuk mengukur kualitas topik yang dihasilkan oleh algoritma LDA. Salah satu metrik yang digunakan adalah *coherence score*, yang mengukur konsistensi kata-kata dalam setiap topik. Selain itu, dilakukan validasi manual untuk memeriksa apakah topik yang ditemukan sesuai dengan konteks data komentar.

Peneliti juga melakukan eksperimen dengan mengubah jumlah topik dan parameter model untuk memastikan hasil terbaik. Misalnya, jumlah topik diubah dari 5 ke 10 atau 15 untuk mengevaluasi bagaimana variasi tersebut memengaruhi kualitas model. Hasil evaluasi digunakan untuk memilih konfigurasi model yang paling sesuai dengan data.

### 3.5. Visualisasi dan Analisis Data

Hasil dari pemodelan topik divisualisasikan menggunakan alat seperti pyLDAvis, yang memungkinkan analisis distribusi topik secara interaktif. Visualisasi ini membantu peneliti untuk memahami bagaimana topik saling berhubungan dan bagaimana komentar tersebar di dalam setiap topik.

Selanjutnya, peneliti menganalisis hasil dengan menafsirkan tema dari setiap topik berdasarkan kata-kata dominan yang dihasilkan oleh model. Topik-topik tersebut kemudian dihubungkan dengan tema besar atau fenomena tertentu yang relevan dengan penelitian.