BAB II
LANDASAN TEORI

2.1 Penelitian Sebelumnya

Bagian ini akan menjelaskan terkait penelitan-penelitian yang telah dilakukan sebelumnya untuk kemudian dijadikan acuan dalam proses penelitian ini. Tabel 2.1 menunjukkan daftar penelitian sebelumnya yang menjadi dasar dalam penelitian ini.

Tabel 2.1 Penelitian sebelumnya

Judul Penelitian	**Metode**	**Penulis**	**Hasil yang Diperoleh**
Analisis Topik Media Sosial Twitter Menggunakan Model Topik *Latent Dirichlet Allocation*	*Latent Dirichlet Allocation*	(Utami, 2017)	Penelitian ini bertujuan untuk melakukan analisis topik dari media sosial Twitter berdasarkan lokasi, yang di dalam kasus ini dibatasi di Kota Bogor pada rentang waktu 1 Februari hingga 7 Februari 2017. Hasil penelitian menunjukkan deskripsi dari kata kunci yang didapat dari hasil pemodelan topik terkait hal yang ramai diperbincangkan pada lokasi dan waktu tertentu.
Text Mining and Topic Modeling on Compendium Papers from Transportation Research Board Annual Meetings	*Latent Dirichlet Allocation*	(Das, Sun, & Dutta, 2016)	Penelitian ini dilakukan untuk mengetahui tren riset penelitian yang terkumpul pada konferensi tahunan dari *Transportation Research Board* selama 7 tahun, yakni mulai tahun 2008 hingga

Judul Penelitian	Metode	Penulis	Hasil yang Diperoleh
			2014. Implementasi penelitian dilakukan dengan menggunakan metode LDA. Hasil yang didapatkan dari penelitian tersebut adalah kumpulan istilah yang paling sering digunakan di dalam judul dan abstrak penelitian, visualisasi pemodelan topik menggunakan *word cloud*, dan visualisasi banyak topik yang muncul berdasarkan tahun menggunakan *area chart*.
Analyzing Research Trends on Drug Safety Using Topic Modeling	*Latent Dirichlet Allocation*	(Zou & Zou, 2018)	Penelitian ini dilakukan untuk mengetahui tren penelitian yang berkaitan dengan *drug safety* berdasarkan judul dan abstrak penelitian yang terdapat di dalam indeks MEDLINE dari tahun 2007 sampai dengan 2016. Hasil penelitian ditunjukkan dengan menampilkan topik riset yang popular dari tahun ke tahun, persebaran topik terhadap kata, dan pengklasteran distribusi topik menggunakan dendogram.

Judul Penelitian	Metode	Penulis	Hasil yang Diperoleh
Penerapan *Information Retrieval* Menggunakan Pemodelan Topik pada Dokumen Skripsi (Studi Kasus Ruang Baca Teknik Informatika UMG)	*Probabilistic Latent Semantic Analysis* (PLSA)	(Waluya, 2017)	Penelitian ini dilakukan untuk menemukan kembali sebuah informasi yang relevan dengan *query* atau kata kunci yang dimasukkan oleh pengguna. PLSA diterapkan sebagai metode untuk pemodelan topik karena dianggap mampu memberikan informasi bagaimana nilai kedekatan antar dokumen, sehingga mampu mengelompokkan dokumen ke dalam topiknya masing-masing. Hasil penelitian menunjukkan bahwa nilai rata-rata *precision* dari aplikasi terhadap *query* adalah 92.8%, *recall* sebesar 91.9%, *accuracy* sebesar 91.7%, dan *F-measure* sebesar 92%.
A Comparative Study of Topic Models for Topic Clustering of Chinese Web News	*Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation,*	(Wu et al., 2010)	Penelitian ini bertujuan untuk melakukan komparasi terhadap tiga metode pemodelan topik dengan menggunakan data dari web berita Cina. Penelitian dilakukan dengan melakukan skenario terhadap nilai *cosine*, dan

Judul Penelitian	Metode	Penulis	Hasil yang Diperoleh
			membandingkannya dengan menggunakan klasterisasi K-center dan menggunakan F-measure untuk menghitung hasil klasterisasi. Hasil penelitian menunjukkan bahwa penggunaan pemodelan topik yang kompleks, seperti PLSA2 dan LDA, kita bisa mendapatkan kelas-kelas topik yang lebih baik daripada menggunakan satu metode pemodelan topik yang sederhana.

Berbeda dari penelitian-penelitian yang telah dilakukan sebelumnya, penelitian ini menggunakan data judul penelitian yang sudah terindeks di dalam Google Scholar. Metode yang digunakan dalam melakukan pemodelan topik ini adalah *Latent Dirichlet Allocation* yang memiliki hasil yang cukup baik dibandingkan dengan metode yang lain. Tahap *preprocessing* juga disesuaikan dengan data yang telah diperoleh. Dalam menampilkan hasil pemodelan topik, akan digunakan visualisasi pyLDAvis dan *word cloud* untuk memudahkan pengguna mengetahui sebaran kata atau frasa di setiap topik.

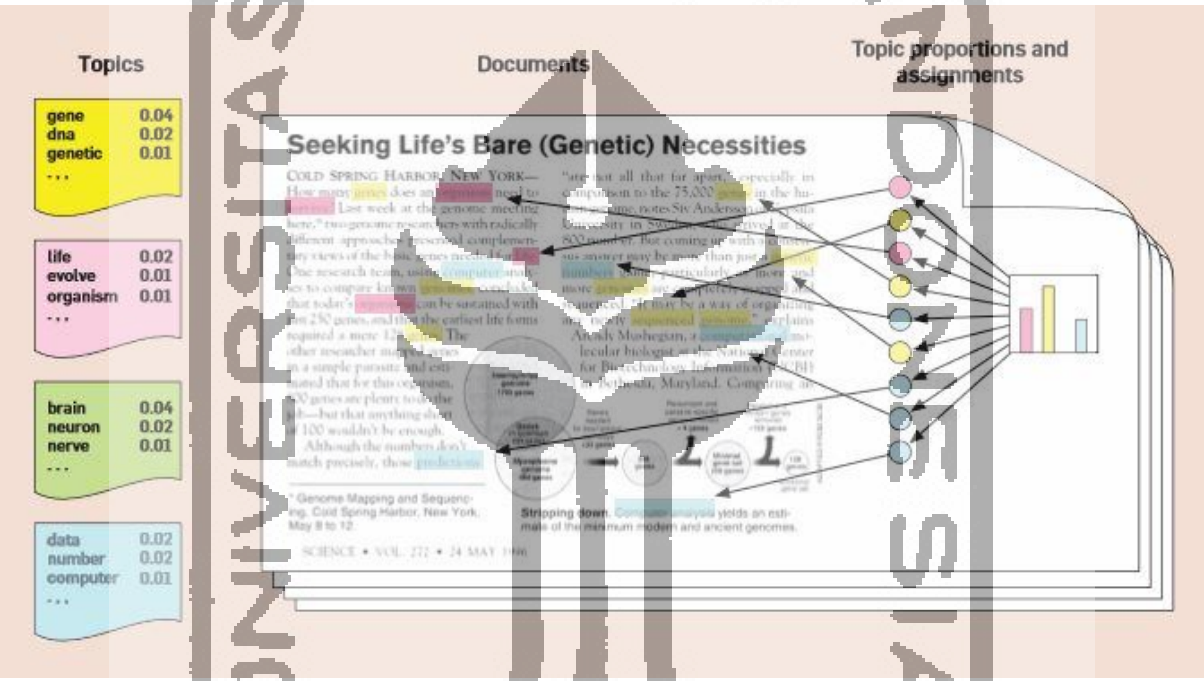
2.2 Dasar Teori

2.2.1 Pemodelan Topik

Pemodelan topik adalah algoritma untuk menemukan tema utama yang berasal dari koleksi dokumen yang besar dan tidak terstruktur. Algoritma tersebut dapat diadaptasi untuk berbagai jenis data (Blei, 2012). Pemodelan topik tersusun dari beberapa komponen yakni, “kata-kata”, “dokumen”, dan “corpora”. Kata merupakan unit dasar dari data diskrit dalam dokumen yang mana setiap kata tersebut memiliki indeks sebagai *identifier*. Dokumen merupakan susunan N kata-kata. Sedangkan *corpora* merupakan bentuk jamak dari korpus.

Korpus yakni kumpulan M dokumen. Secara sederhana, setiap dokumen dalam korpus mempunyai topik-topik yang dibahas berdasarkan kata-kata yang terkandung di dalamnya (Blei, Edu, et al., 2003).

Algoritma pemodelan topik tidak memerlukan penjelasan atau pelabelan dari dokumen sebelumnya, karena topik dapat muncul dari analisis teks asli. Pemodelan topik memungkinkan kita untuk mengatur dan meringkas arsip elektronik pada skala yang mungkin tidak bisa dilakukan oleh anotasi manusia (Blei, 2012). Secara konsep dan teori, visualisasi cara kerja pemodelan topik dapat dilihat pada Gambar 2.1.



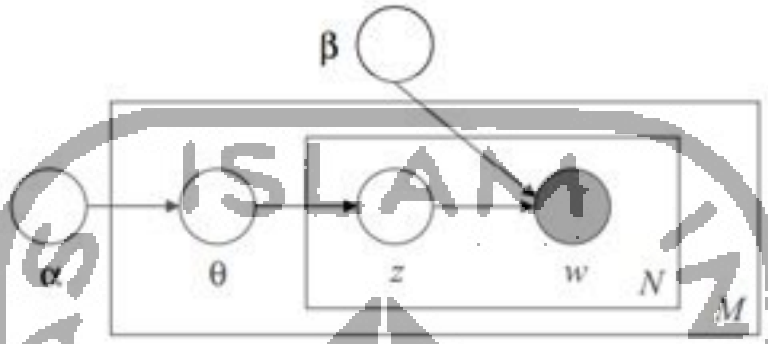
Gambar 2.1 Konsep pemodelan topik
(Blei, 2012)

2.2.2 *Latent Dirichlet Allocation* (LDA)

Latent Dirichlet Allocation (LDA) merupakan salah satu metode yang digunakan untuk melakukan pemodelan topik pada data berbentuk teks. Ide yang mendasar dari LDA adalah bahwa setiap dokumen direpresentasikan sebagai campuran dari topik-topik yang tersembunyi dan belum diketahui, di mana setiap topik terdiri dari distribusi berbagai kata (Blei, Ng, & Jordan, 2003). Model LDA melakukan proses generatif untuk setiap dokumen dalam korpus dengan tahahapan sebagai berikut (Blei, Ng, et al., 2003):

1. Memilih topik secara acak berdasarkan distribusi topik-topik pada setiap dokumen.

2. Memilih kata berdasarkan distribusi berbagai kata yang berkaitan dengan topik terpilih.
3. Mengulangi tahapan 1 dan 2 pada seluruh dokumen yang ada.



Gambar 2.2 Representasi model LDA

Model LDA direpresentasikan sebagai model grafis probabilistik yang dibagi ke dalam 3 tingkatan. Parameter α dan β adalah parameter dari tingkatan korpus, yang diasumsikan menjadi sampel dalam proses menghasilkan korpus. Parameter α digunakan untuk menentukan distribusi topik dalam dokumen. Hal ini menunjukkan bahwa semakin besar nilai α, maka semakin banyak topik yang akan dibahas. Sedangkan parameter β digunakan untuk menentukan distribusi kata pada setiap topik. Hal ini menunjukkan semakin kecil nilai β maka semakin spesifik kata-kata yang berada pada topik. Variabel θ_m adalah variabel dari tingkatan dokumen M, di mana M menunjukkan adanya perulangan variabel sebanyak M kali pada setiap dokumen. Variabel θ_m merepresentasikan distribusi topik pada dokumen tertentu. Semakin kecil nilai θ, maka semakin spesifik topik yang berada pada dokumen. Variabel z_n dan w_n adalah variabel dari tingkatan kata dalam dokumen N. Sama halnya dengan M, N juga menunjukkan adanya perulangan pada variabel di dalamnya sebanyak N kali, namun pada setiap kata. Variabel z adalah variabel yang merepresentasikan topik dari kata tertentu pada sebuah dokumen, sedangkan variabel w adalah variabel yang merepresentasikan keterkaitan kata dengan topik tertentu yang berada pada sebuah dokumen (Blei, Ng, et al., 2003).

2.2.3 *Topic Coherence*

Salah satu cara dalam menentukan cacah topik yang ideal yang dihasilkan dalam proses pemodelan topik adalah dengan menggunakan *topic coherence*. Penggunaan *topic coherence*

dilakukan untuk memberikan rekomendasi banyak topik yang akan dibentuk setelah model dibentuk.

Topic coherence—yang bermakna koherensi semantik—di mana kualitas penilaiannya didasarkan pada bagaimana manusia memahami semantik kata, dan tidak bisa diukur dengan pengukuran statistik berbasis model (Newman, Bonilla, & Buntine, n.d.). Namun penelitian terkini telah mendemonstrasikan bahwa ternyata adalah suatu hal yang mungkin untuk mengukur *topic coherence* dengan akurasi yang mendekati sebagaimana apa yang dapat dilakukan oleh manusia berdasarkan nilai *pointwise mutual information* (PMI) (Newman, Lau, Grieser, & Baldwin, 2010).

2.2.4 *Preprocessing*

Preprocessing merupakan tahapan untuk mempersiapkan data setelah sebelumnya dikumpulkan. Dengan *preprocessing*, kita dapat menyeragamkan standar kualitas data yang akan diproses. Seringkali data yang didapat dirasa tidak layak untuk diproses, seperti data yang berganda, nilai yang kosong atau hilang, adanya kesalahan dalam pengisian data, ketidaksesuaian pengisian data, atau bentuk data yang tidak ternormalisasi. Hasil dari *preprocessing* akan berupa nilai numerik sehingga dapat dijadikan sebagai sumber data yang dapat diolah lebih lanjut (Socrates, Akbar, Akbar, Arifin, & Herumurti, 2017).

Sebelum melakukan proses pengolahan bahasa alami, tahap *preprocessing* adalah tahap yang perlu dilakukan dan diperhatikan untuk menghilangkan—atau setidaknya mengurangi—kondisi data yang tidak konsisten dan tidak beraturan. Tahap *preprocessing* memiliki pengaruh yang cukup signifikan dalam memberikan hasil setelah dilakukannya pemrosesan atau pengolahan data teks. Tahap *preprocessing* yang dilakukan dengan baik akan mendukung hasil pengolahan yang lebih baik juga. Beberapa hal yang dapat dilakukan di dalam tahap ini di antaranya adalah menghilangkan angka dan simbol, *case folding*, *stemming*, dan menghapus *stopwords.*

2.2.5 Visualisasi *Dashboard*

Dashboard adalah kumpulan komponen informatif yang dikelompokkan bersama. Komponen informatif itu dapat berupa laporan analisis, grafik, *scorecard* atau kombinasi dari semuanya (Rizq, Trisna, Ahmadi, & Suardika, n.d.).

Dashboard dapat digunakan untuk melihat dan memahami hasil dari pengolahan data. *Dashboard* dapat dijadikan sebuah solusi untuk menjembatani pemahaman pengembang atau

peneliti dengan pengguna yang mungkin terjadi. Dengan adanya *dashboard*, maka pengguna akan mudah dalam memahami dan menginterpretasi hasil dari pengolahan data tanpa harus memahami konsep pengolahannya secara teknis dan mendalam. *Dashboard* yang didesain dengan baik akan mendukung pemahaman pengguna yang lebih baik pula terhadap informasi yang disampaikan.