BAB II
TINJAUAN PUSTAKA

A. Peneliti Terdahulu

Tujuan dari penelitian sebelumnya adalah untuk mengumpulkan data referensi dan perbandungan. Selain itu untuk menghindari anggapan bahwa penelitian ini sama dengan penelitian lainnya. Sebagai konsekuensi, peneliti memasukan temuan penelitiannya yang di tampilkan dalam tabel 2.1 di bawah ini dalam tinjauan pustaka

Tabel 2.1 Peneliti Terdahulu

No	Nama	Judul	Hasil
1	(Rahmawati et al., 2021)	Analisis topik konten channel YouTube K-pop Indonesia menggunakan *Latent Dirichlet Allocation*	Nilai coheren dari 5 topik 0,06294, Keywords, unboxing, album, reaction pada topik ke 5
2	(Matira et al., 2023)	Pemodelan Topik pada Judul Berita Online Detikcom Menggunakan *Latent Dirichlet Allocation*	Nilai coheren dari 3 topik 0,7586 Keywords, topik 1 bencana alam, topik 2 korban perang ukrain, topik 3 korupsi.
3	(Nazila & Utari, 2023)	Pemodelan Topik Keluhan Masyarakat Pasca Pandemi Menggunakan *Metode Latent Dirichlet Allocation* (LDA)	*Akurasi nilai topik 21% keywords, jalan, layan, rusak, godean, proyek, online, jl, dukcapil, adu, sleman.*
4	(Cxallı & Cxallı, 2023)	*Understanding Airline Passengers during Covid-19 Outbreak to Improve Service Quality: Topic Modeling Approach to Complaints with Latent Dirichlet Allocation Algorithm*	*Topics accuracy, 92.14% pegasus airlines keywords, family member, website. Topics (accuracy: 87.86%) turkish airline keywords, refund cancellation, baggage.*
5	(Breuninger et al., 2021)	*Associations between habitual diet, metabolic disease, and the gut*	Topik akurasi 74,76% dari topik genus *keywords, bacteroides, dialister, parabacteroides.*

No	Nama	Judul	Hasil
		microbiota using latent Dirichlet allocation	
6	(Zou et al., 2022)	*Public perceptions of digital fashion: An analysis of sentiment and Latent Dirichlet Allocation topic modeling*	Nilai akurasi topik dua 82,5% *keywords, fashion, digital, daily, global, stories, future, metaverse, virtual, nfts, week,*
7	(Zulhanif et al., 2017)	Aplikasi *Latent Dirichlet Allocation* (Lda) Pada *Clustering* Data Teks	Nilai akurasi pada topik empat 4,3% *keywords*, alsyariier, pict, aliandofan, latepost, java, travel, ztyspui, west, ogartyoc.
8	(Luvian chisni chilmi, 2021)	*Latent Dirichlet Allocation* (LDA) Untuk Mengetahui Topik Pembicaraan Warganet *Twitter* Tentang Omnibus Law	Nilai akurasi pada topik dua 17,8% *keywords*, cekal koorona, korona, ikut demo, presiden jokowi, demo, pemerintah, cipta kerja, februari
9	(Dikiyanti et al., 2021)	*Sentiment analysis and topic modeling of BPJS Kesehatan based on twitter crawling data using Indonesian Sentiment Lexicon and Latent Dirichlet Allocation algorithm*	Sentimen nilai akurasi 61,7% positive, dan 38,3% negativ. Pada topik modeling BPJS kesehatan
10	(Putra & Kusumawardani, 2017)	Analisis Topik Informasi Publik Media Sosial Di Surabaya Menggunakan Pemodelan *Latent Dirichlet Allocation*	*Coherence score* 5 topik 288,16 *keywords* topik 2, padat, arah, macet, jalur, imbas, tol, via, surabaya.

B. Landasan Teori

Klasterisasi merupakan pembagian kumpulan item data ke dalam kumpulan komponen yang dikenal sebagai cluster dikenal sebagai pengelompokan. *Clustering* merupakan salah satu teknik data mining yang digunakan untuk mendapatkan kelompok-kelompok dari obyek-obyek yang mempunyai karakteristik yang umum di data yang cukup besar. Tujuan utama dari metode *clustering* adalah pengelompokan sejumlah data/obyek

ke dalam cluster (group) sehingga dalam setiap cluster akan berisi data yang semirip mungkin (Nugraha et al., 2014). Oleh karena itu clustering banyak di gunakan sebagai metode untuk pengelompokan data contohnya, mengelompokan data costumer dan lain lain sebagainya.

1. Otomotif

Pengertian otomotif adalah jika dilihat dari fungsi kata otomotif yang berkedudukan sebagai kata sifat, otomotif merupakan sesuatu yang berhubungan dengan alat yang dapat berputar atau bergerak dengan sendirinya. Otomotif biasanya akan dikaitkan dengan motor atau mesin yang dapat menggerakan benda yang lebih besar daripada mesin/motor penggerak tersebut. Otomotif juga mempunyai kaitan yang sangat erat dengan dunia industri dan transportasi di mana kedua bidang tersebut pada umumnya akan digunakan tenaga mesin dan mobi. Hal-hal yang mencakup kedalam bidang otomotif antara lain, perencanaan, pengembangan, produksi, dan perawatan (Aryawan et al., 2019).

2. YouTube

YouTube adalah situs web yang dirancang untuk memungkinkan orang melihat dan menikmati video, serta berbagi video yang mereka miliki. Anda dapat menemukan video musik, film pendek, serial TV, trailer film, video pendidikan, dan banyak lagi di *youtube*, yang semuanya dibuat oleh orang yang berbeda (Wiryany & Pratami, 2019). Saat ini *Youtube* menjadi situs online video provider paling dominan di Amerika serikat, bahkan dunia, dengan menguasai 43% pasar. Diperkirakan 20 Jam durasi video di

upload ke *YouTube* setiap menitnya dengan 6 miliar views per hari. *YouTube* kini telah menjadi berbagai macam kebutuhan dari penggunanya, fiturfitur yang ditawarkan dengan kemajuan teknologi *YouTube* saat ini sangat membantu dari berbagai aspek kebutuhan yang dibutuhkan sang pengguna (Thanissaro & Kulupana, 2015).

3. Scraping

Scraping adalah teknik untuk mendapatkan informasi dari *website* secara otomatis tanpa harus menyalinnya secara manual. Tujuan dari web *scraper* adalah untuk mencari informasi tententu dan kemudian mengumpulkannya kedalam format yang baru. *Web scraping* berfokus dalam mendapatkan data dengan cara pengambilan dan ekstraksi (A. Yani et al., 2019). Teknik-teknik web scraping diantaranya adalah, penggunaan *api*, merupakan (*Application Programming Interface*) untuk mengakses data mereka dengan cara yang terstruktur dan terdokumentasi. API ini memungkinkan untuk mengambil data dengan lebih mudah dan legal. Dari data ini kemudian yang akan di kelompokan dan lakukan proses topik modeling menggunakan LDA.

4. Text Mining

Salah satu jenis data mining yang menggunakan bahasa atau tulisan dan menarik informasinya dari dokumen disebut text mining. Untuk menganalisis data yang dikumpulkan dan menentukan hubungannya dengan dokumen lain, tujuan dari *text mining* adalah untuk mengidentifikasi kata-kata yang dapat mewakili sebagian besar isi dokumen. Teknik-teknik *text*

mining ini digunakan untuk format dokumen yang tidak terstruktur, ambigu, dan sulit diuraikan sebagai sumber data. Baik dokumen statis maupun dinamis dapat digunakan dalam proses *text mining* ini (Rakhmawati et al., 2021).

5. Google Colaboratory

Google Research menciptakan *Google Collaboratory*, terkadang disebut sebagai *Colab*, sebuah platform berbasis cloud untuk pengajaran dan penelitian pembelajaran mesin yang dibangun di atas *Jupyter Notebooks.* (Gustrio, 2015). Dalam studi ini, *Google Collaboratory* dipilih karena sifatnya yang *open-source* dan integrasinya yang mulus dengan *Google Drive*, menjadikannya *platform* yang ramah pengguna untuk menulis, menyimpan, dan berbagi program.

6. Python

Guido van Rossum menciptakan bahasa pemrograman tingkat tinggi *python* pada tahun 1989, dan awalnya dirilis pada tahun 1991. *Python* adalah bahasa pemrograman untuk mempermudah programmer melakukan pekerjaan mereka dengan cepat. *Python* dapat digunakan untuk menulis skrip dan aplikasi mandiri. *Python* bisa juga untuk menawarkan sejumlah manfaat, termasuk pemrograman *python* dapat menangani pemrograman yang rumit dan memungkinkan pemrograman grafis. Ini juga platform netral, artinya aplikasi yang ditulis dengan *python* dapat berjalan di sistem operasi apa pun selama ada platform *python.* Dibandingkan dengan *C++* dan *Java*, *python* secara signifikan lebih cepat dan lebih pendek untuk

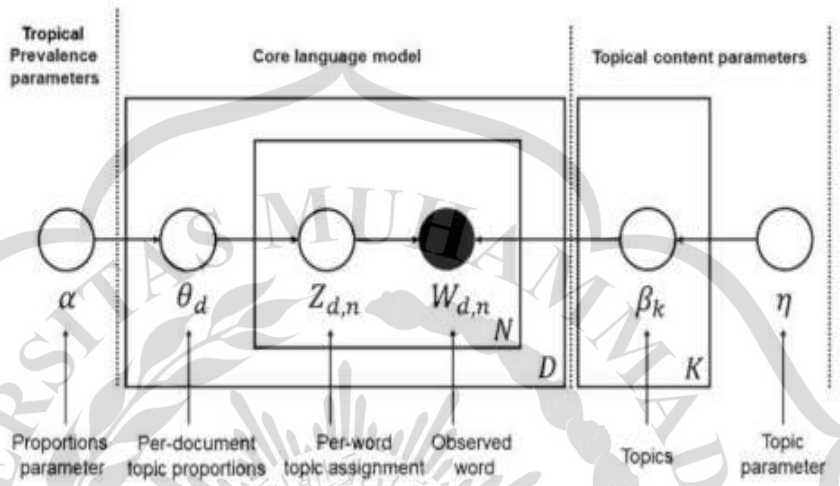
digunakan (Phandany et al., 2022). Python adalah bahasa pemrograman komputer yang sering digunakan untuk membangun situs, software/aplikasi, membuat tugas menjadi otomatis dan menganalisis data secara cepat dan detail. Bahasa pemrograman *python* termasuk bahasa tujuan umum. Maksudnya adalah bisa untuk digunakan di berbagai program berbeda, bukan khusus untuk masalah tertentu saja. Python ini biasa di gunakan dalam pengembangan situs dan perangkat lunak, membuat analisis data, visualisasi data dan otomatisasi tugas. Karena sifatnya yang relatif mudah dipelajari, bahasa pemrograman ini digunakan secara luas oleh banyak orang seperti ilmuwan dan akuntan untuk melakukan tugas harian mereka. Misalnya, dalam mengatur keuangan dan mengelola data pada suatu dokumen.

7. Metode *Lattent Dirichlet Allocation*

Lattent Dirichlet Allocation dalah kumpulan dokumen dan beberapa parameter. Outputnya adalah model yang terdiri dari bobot yang dapat dinormalisasi menjadi probabilitas. Probabilitas ini terdiri dari dua jenis: Yang pertama bahwa probabilitas dokumen tertentu menghasilkan topik tertentu pada suatu posisi, dan kedua yaitu probabilitas dari topik menghasilkan kata tertentu dari kumpulan kosakata (Campbell et al., 2014). Penelitian ini menerapkan metode topic modelling LDA yang bertujuan untuk melakukan analisis tren topik yang akan dihasilkan dan divisualisasikan sehingga akan lebih informatif dan mudah dipahami oleh pengguna. kemudian hasil dari penelitian ini menunjukan topik model untuk

mengelompokkan dokumen dan menemukan dokumen yang serupa (Sahria & Fudholi, 2017).

Rumus dan representasi LDA ditunjukan sebagai berikut :



Gambar 2.1 Diagram Model LDA (Zou et al., 2022)

Model LDA dapat direpresentasikan sebagai model grafis probabilistik seperti pada Gambar 3.2 menjelaskan, ada tiga tingkat sebagai representasi LDA. Parameter α dan β sebagai corpus level parameter, diasumsikan disampel sekali dalam proses menghasilkan sebuah corpus. Variabel θd adalah adalah variabel pada tingkat dokumen, disampel sekali per dokumen. Akhirnya, variabel Z_{dn} dan W_{dn} yang merupakan variabel pada tingkat.Perhitungan probabilitas dari sebuah corpus berdasarkan persamaan LDA yang telah dijelaskan dapat dilihat bahwa pada notasi β mendeskripsikan topik, dimana pada setiap β merupakan distribusi dari sejumlah kata. Pada Variabel θd adalah variabel level dokumen dengan satu kali sampel per dokumen yang merepresentasikan proporsi topik untuk

dokumen ke d. Pada notasi Z_{dn} dan W_{dn} merupakan representasi variabel di level kata dengan satu kali sampel untuk masing-masing kata pada setiap dokumen.