PEMODELAN TOPIK OPINI MASYARAKAT PENGGUNA TWITTER TERHADAP PT KERETA API INDONESIA (PERSERO) MENGGUNAKAN *LATENT DIRICHLET ALLOCATION*

SKRIPSI

UIN SUNAN AMPEL
S U R A B A Y A

Disusun Oleh

YASIRAH REZQITA AISYAH YASMIN

H92219062

PROGRAM STUDI MATEMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SUNAN AMPEL

SURABAYA

2023

PERNYATAAN KEASLIAN

Saya yang bertanda tangan di bawah ini,

Nama	:	YASIRAH REZQITA AISYAH YASMIN
NIM	:	H92219062
Program Studi	:	Matematika
Angkatan	:	2019

Menyatakan bahwa saya tidak melakukan plagiat dalam penulisan skripsi saya yang berjudul "Pemodelan Topik Opini Masyarakat Pengguna Twitter Terhadap PT Kereta Api Indonesia (Persero) Menggunakan *Latent Dirichlet Allocation*". Apabila suatu saat nanti terbukti saya melakukan tindakan plagiat, maka saya bersedia menerima sanksi yang telah ditetapkan.

Demikian pernyataan keaslian ini saya buat dengan sebenar-benarnya.

Surabaya, 10 Juli 2023

Yang menyatakan,

YASIRAH REZQITA AISYAH YASMIN

NIM. H92219062

LEMBAR PERSETUJUAN PEMBIMBING

Skripsi oleh

Nama	:	YASIRAH REZQITA AISYAH YASMIN
NIM	:	H92219062
Judul Skripsi	:	Pemodelan Topik Opini Masyarakat Pengguna Twitter Terhadap PT Kereta Api Indonesia (Persero) Menggunakan *Latent Dirichlet Allocation*

telah diperiksa dan disetujui untuk diujikan.

Pembimbing I	Pembimbing II
Nurrissaidah Ulinnuha, M.Kom	Lutfi Hakim, M.Ag
NIP. 199011022014032004	NIP. 197312252006041001

Mengetahui,
Ketua Program Studi Matematika
UIN Sunan Ampel Surabaya

Yuniar Farida, MT
NIP. 197905272014032002

PENGESAHAN TIM PENGUJI SKRIPSI

Skripsi oleh

Nama	:	YASIRAH REZQITA AISYAH YASMIN
NIM	:	H92219062
Judul Skripsi	:	Pemodelan Topik Opini Masyarakat Pengguna Twitter Terhadap PT Kereta Api Indonesia (Persero) Menggunakan *Latent Dirichlet Allocation*

Telah dipertahankan di depan Tim Penguji
pada tanggal 03 Juli 2023

Mengesahkan,
Tim Penguji

Penguji I

Dr. Dian Candra Rini Novitasari, M.Kom
NIP. 198511242014032001

Penguji II

Aris Fanani, M.Kom
NIP. 198701272014031002

Penguji III

Nurrissaidah Ulinnuha, M.Kom
NIP. 199011022014032004

Penguji IV

Dr. Lutfi Hakim, M.Ag
NIP. 197312252006041001

Mengetahui,
Dekan Fakultas Sains dan Teknologi
UIN Sunan Ampel Surabaya

Dr. A. Saepul Hamdani, M.Pd
NIP. 196507312000031002

UIN SUNAN AMPEL
S U R A B A Y A

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL SURABAYA
PERPUSTAKAAN
Jl. Jend. A. Yani 117 Surabaya 60237 Telp. 031-8431972 Fax.031-8413300
E-Mail: perpus@uinsby.ac.id

LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI
KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika UIN Sunan Ampel Surabaya, yang bertanda tangan di bawah ini, saya:

Nama : YASIRAH REZQITA AISYAH YASMIN
NIM : H92219062
Fakultas/Jurusan : FAKULTAS SAINS DAN TEKNOLOGI / MATEMATIKA
E-mail address : yasirahrqis@gmail.com

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Perpustakaan UIN Sunan Ampel Surabaya, Hak Bebas Royalti Non-Eksklusif atas karya ilmiah :
☑ Skripsi ☐ Tesis ☐ Disertasi ☐ Lain-lain (.....)
yang berjudul :

PEMODELAN TOPIK OPINI MASYARAKAT PENGGUNA TWITTER TERHADAP PT KERETA API INDONESIA (PERSERO) MENGGUNAKAN LATENT DIRICHLET ALLOCATION

beserta perangkat yang diperlukan (bila ada). Dengan Hak Bebas Royalti Non-Ekslusif ini Perpustakaan UIN Sunan Ampel Surabaya berhak menyimpan, mengalih-media/format-kan, mengelolanya dalam bentuk pangkalan data (database), mendistribusikannya, dan menampilkan/mempublikasikannya di Internet atau media lain secara ***fulltext*** untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan atau penerbit yang bersangkutan.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Perpustakaan UIN Sunan Ampel Surabaya, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini yang saya buat dengan sebenarnya.

Surabaya, 10 JULI 2023

Penulis

(YASIRAH REZQITA A.Y)
nama terang dan tanda tangan

ABSTRAK

Pemodelan Topik Opini Masyarakat Pengguna Twitter Terhadap PT Kereta Api Indonesia (Persero) Menggunakan *Latent Dirichlet Allocation*

Kereta api merupakan transportasi umum yang diminati masyarakat karena harganya yang murah dengan waktu perjalanan yang efisien. Di indonesia kereta api dikelola oleh PT Kereta Indonesia (Persero). Dalam menjalani kehidupan perusahaan, PT Kereta Api Indonesia (Persero) memiliki budaya perusahaan yang biasa disebut "AKHLAK". Penyampaian kebudayaan tersebut kepada masyarakat salah satunya dengan media sosial. Twitter menjadi akun resmi media sosial dengan followers terbanyak dibandingkan akun media sosial lainnya. Oleh karena itu, perlu dilakukannya pemodelan topik untuk mengetahui topik apa saja yang dibicarakan masyarakat pengguna twitter terhadap PT Kereta Indonesia Power. Pemodelan topik dapat dilakukan menggunakan metode Latent Dirichlet Allocation (LDA). LDA merupakan model probabilistik untuk pemodelan topik berdasarkan data tekstual untuk mendapatkan informasi yang berupa model topik. Pada penelitian ini dilakukan pemodelan topik LDA menggunakan 2 macam pembobotan kata yakni *term frequency* (TF) dan *Term frequency inverse document frequency* (TF-IDF). Didapatkan model LDA lebih baik dengan 4 topik menggunakan pembobotan TF-IDF dengan nilai koherensi sebesar 0.735. Sementara pada pembobotan TF nilai koherensi terbesar hanya 0.69 dengan jumlah topik sebanyak 10.

Kata kunci: *Latent Dirichlet Allocation*, Nilai koherensi, Pemodelan Topik, *Text Mining*

ABSTRACT

Topic Modeling the Opinion of the Twitter User Community Against PT Kereta Api Indonesia (Persero) Using Latent Dirichlet Allocation

The train is public transportation that is of interest to the public because of its low price and efficient travel time. In Indonesia, trains are managed by PT Kereta Indonesia (Persero). In living the life of the company, PT Kereta Api Indonesia (Persero) has a corporate culture commonly called ”AKHLAK”. One way to convey this culture to the public is through social media. Twitter is the official social media account with the most followers compared to other social media accounts. Therefore, it is necessary to do topic modeling to find out what topics are discussed by the Twitter user community about PT Kereta Indonesia Power. Topic modeling can be done using the Latent Dirichlet Allocation (LDA) method. LDA is a probabilistic model for modeling topics based on textual data to obtain information in the form of topic models. In this study, LDA topic modeling was carried out using 2 kinds of word weighting, namely term frequency (TF) and Term frequency inverse document frequency (TF-IDF). The LDA model is better with 3 topics using TF-IDF weighting with a coherence value of 0.6970. Meanwhile, for the TF weighting, the highest coherence value is only 0.4958 with a total of 2 topics.

Keywords: Latent Dirichlet Allocation, Coherence Score, Topic Modeling, Text Mining

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Kereta api merupakan transportasi umum yang diminati masyarakat karena harganya yang murah dengan waktu perjalanan yang efisien. Di indonesia kereta api dikelola oleh badan usaha milik negara (BUMN) bernama PT Kereta Indonesia (Persero). Tercatat oleh Badan Pusat Statistik (BPS) jumlah penumpang kumulatif pada bulan Januari 2022 hingga bulan September 2022 mencapai 192,8 juta penumpang. Jumlah tersebut mengalami kenaikan sebanyak 85,76% dibandingkan dengan jumlah penumpang pada tahun 2021 (Mutia, 2022). Kenaikan tersebut menunjukkan Bahwa, PT Kereta Api Indonesia (Persero) secara konsisten berinovasi dalam kemajuan perkeretaapian Indonesia (CNN Indonesia, 2017).

Dalam menjalani kehidupan perusahaan, PT Kereta Api Indonesia (Persero) memiliki budaya perusahaan yang biasa disebut "AKHLAK" yang merupakan kepanjangan dari Amanah, Kompeten, Harmonis, Loyal, Adaptif, dan Kolaboratif (BUMN, 2022). Kebudayaan tersebut merupakan salah satu bentuk profesionalitas PT Kereta Api Indonesia kepada pengguna jasanya. Sikap profesionalitas tercerminkan dalam al-Qur'an surat Al-Isra ayat 84 yang berbunyi:

قُلْ كُلٌّ يَّعْمَلُ عَلٰى شَاكِلَتِهٖۗ فَرَبُّكُمْ اَعْلَمُ بِمَنْ هُوَ اَهْدٰى سَبِيْلًا ۧ

artinya: Katakanlah (Muhammad), "Setiap orang berbuat sesuai dengan

pembawaannya masing-masing." Maka Tuhanmu lebih mengetahui siapa yang lebih benar jalannya. (QS. Al-Isra: 84)

Inti dari surah Al-Isra ayat 84 adalah Allah memerintahkan kepada hamba-Nya untuk melakukan pekerjaannya dengan penuh ketekunan, Jika seseorang bekerja dengan penuh ketekunan akan menghasilkan hasil yang optimal. Dalam penyampaian kebudayaan "AKHLAK" tersebut kepada masyarakat perlu dibentuknya *company profile*. *Company profile* adalah sebuah cara dalam menyampaikan pesan perusahaan kepada masyarakat (Muhammad et al., 2018). Banyak berbagai macam media dalam menyampaikan *company profile* yang baik, salah satunya yakni media sosial.

Media sosial merupakan sebuah wadah bertukar informasi secara berkelanjutan dengan jangkauan yang sangat luas dan waktu yang singkat. Postingan media sosial dapat menjadi sumber data untuk melihat respons terhadap situasi yang signifikan (Weidner et al., 2021). Oleh karena itu, banyak perusahaan menggunakan media sosial resmi untuk berinteraksi kepada pelanggan. Seperti halnya PT Kereta Api Indonesia (Persero) memiliki akun official di berbagai macam sosial media seperti facebook, instagram, twitter, youtube, maupun tiktok. Instagram memiliki pengikut sebanyak 911 ribu, tiktok memiliki pengikut sekitar 322 ribu, dan facebook memiliki 232 ribu pengikut. Dari berbagai macam akun resmi media sosial yang dimiliki PT Kereta Api Indonesia (Persero), twitter menjadi akun resmi media sosial dengan followers terbanyak dibandingkan akun media sosial lainnya.

Akun resmi twitter PT Kereta Api Indonesia (persero) dengan nama pengguna "@KAI121" memiliki jumlah pengikut yang paling banyak yakni lebih dari satu juta akun pengikut. Lebih dari 50 tweets setiap harinya dengan

percakapan mention sebesar 99% (Twitter, 2022). Hal tersebut menandakan bahwa banyak pengguna kereta yang memanfaatkan media sosial twitter sebagai sarana komunikasi langsung terhadap PT Kereta Api Indonesia (Persero).

Twitter merupakan media sosial kedua paling banyak dikunjungi di seluruh dunia, sekitar 192 juta pengguna aktif dengan 500 juta tweet dibagikan setiap hari (Thakur, 2022). Karena hanya terdiri dari 280 karakter twitter dianggap dapat menangkap informasi singkat dan padat (*microblogging*) (Fikriyah and Sibaroni, 2022). Karena twitter merupakan sosial media berdasarkan *microblogging* dengan algoritma yang unik, dimana *tweet* pengguna lain dapat muncul dalam beranda tanpa harus saling mengikuti. Hal tersebut dapat membuat suatu opini dalam *tweet* secara mudah terekspos masyarakat atau viral. Oleh karena itu, tidak sedikit masyarakat Indonesia yang menggunakan twitter sebagai wadah dalam menyampaikan opini terhadap kinerja suatu perusahaan.

Penyampaian opini atau pendapat merupakan salah satu cara langkah dalam bermusyawarah untuk menyelesaikan masalah. Dalam penyampaian pendapat, perlu adanya tata krama di dalamnya. Yakni seperti berbicara yang baik dan tidak menghakimi seseorang. Cara penyampaian pendapat dalam bermusyawarah untuk memecahkan masalah tersebut diajarkan di dalam islam. Allah sudah mengatur hamba-Nya untuk mengutarakan opini pribadi dengan cara yang baik, hal tersebut tercantum dalam surah Al-Imran ayat 159 yang berbunyi:

فَبِمَا رَحْمَةٍ مِّنَ ٱللَّهِ لِنتَ لَهُمْ ۖ وَلَوْ كُنتَ فَظًّا غَلِيظَ ٱلْقَلْبِ لَٱنفَضُّوا۟ مِنْ حَوْلِكَ ۖ فَٱعْفُ عَنْهُمْ وَٱسْتَغْفِرْ لَهُمْ وَشَاوِرْهُمْ
فِى ٱلْأَمْرِ ۖ فَإِذَا عَزَمْتَ فَتَوَكَّلْ عَلَى ٱللَّهِ ۚ إِنَّ ٱللَّهَ يُحِبُّ ٱلْمُتَوَكِّلِينَ ﴿١٥٩﴾

artinya: Maka disebabkan rahmat dari Allah-lah kamu berlaku lemah lembut terhadap mereka. Sekiranya kamu bersikap keras lagi berhati kasar, tentulah

mereka menjauhkan diri dari sekelilingmu. Karena itu maafkanlah mereka, mohonkanlah ampun bagi mereka, dan bermusyawarahlah dengan mereka dalam urusan itu. Kemudian apabila kamu telah membulatkan tekad, maka bertawakallah kepada Allah. Sesungguhnya Allah menyukai orang-orang yang bertawakal kepada-Nya (QS. Al-Imran: 159).

Inti dari surah Al-Imran ayat 159, Allah memerintahkan hamba-Nya untuk berperilaku lemah lembut dan bermusyawarah. Hal tersebut dapat dimaknai bahwa kita sebagai manusia seharusnya menjaga sikap untuk tetap berperilaku lemah lembut di segala keadaan. Sikap lemah lembut tersebut dapat mengatasi masalah dengan kepala dingin sehingga menyampaikan pendapat dapat dipahami dan diterima oleh masyarakat. Maka dari itu, tidak heran untuk manusia saling mengutarakan opini pribadi terhadap sesuatu baik disampaikan secara langsung maupun disampaikan pada media tertentu. Salah satunya opini publik terhadap PT Kereta Api Indonesia (Persero) yang tertuang di dalam media sosial twitter.

Oleh karena itu, twitter dapat menyimpan banyak data yang terdapat di dalam *tweets* para pengguna (Pratama et al., 2019). Walaupun terdapat banyak data di dalam *tweets*, data tersebut masih berupa data yang tidak terstruktur, artinya data tersebut sulit untuk dipahami (Curiskis et al., 2020). Sehingga data yang tidak terstruktur tersebut perlu dilakukannya analisis sentimen agar informasi dapat dipahami. Berdasarkan hal tersebut perlu dilakukannya analisis sentimen untuk mengetahui topik topik mana saja yang dibahas pada twitter tentang opini pada PT Kereta Api Indonesia (Persero).

Teks mining merupakan suatu proses untuk mencari informasi dari beberapa dokumen yang berbentuk teks (Ranjbari et al., 2021). Algoritma pada text mining mampu mengidentifikasi data-data semi terstruktur dan tidak terstruktur seperti

email, dokumen, hingga komentar di media sosial (Thakur and Kumar, 2022). Salah satu teknik dari *text mining* untuk menemukan pola dari dokumen tekstual yang masih utuh adalah pemodelan topik (Mohammadi and Karami, 2022). Salah satu metode pemodelan topik terpopuler saat ini adalah LDA. LDA adalah model probabilistik untuk pemodelan topik berdasarkan data tekstual untuk mendapatkan informasi yang berupa model topik (Natalia et al., 2021). LDA dapat menghasilkan sistem yang menentukan kemungkinan beberapa kelompok topik yang terdiri dari beberapa kata. Pada penelitian sebelumnya LDA dapat mengelompokkan lebih dari 5000 calon mitra kerja menjadi 10 Sektor (Kang et al., 2019). Hasil penelitian (Sahria and Fudholi, 2017) menyatakan bahwa LDA mampu mengelompokan penelitian pada bidang kesehatan di Indonesia menghasilkan 3 topik dengan 2 topik yang dominan, lalu hasil tersebut disampaikan kepada responden peneliti, tenaga kesehatan, dan akademisi sebanyak 5,9% mengatakan baik dan 94,1% mengatakan sangat baik. Penelitian sebelumnya terkait pemodelan topik LDA menggunakan data twitter dengan kata pencarian ”Indonesia” menghasilkan 10 topik terbaik dari jumlah data sebanyak 9094 data tweet (Negara and Triadi, 2021).

Pada penelitian yang dilakukan oleh (Brito et al., 2021) mendapatkan hasil penelitian yang menyatakan bahwa algoritma LDA lebih baik digunakan dalam sistem dengan arsitektur yang kompleks hingga dapat mengambil informasi yang lebih luas. Penelitian lainnya membandingkan model algoritma LDA dengan LSA menggunakan kumpulan sepuluh ribu berita acak BBC menghasilkan model LDA lebih baik daripada LSA. Karena LDA dapat menggeneralisasi ke dokumen baru dengan mudah (Kalepalli et al., 2020). Penelitian yang serupa pada pemodelan topik E-Book dengan total 300 buku serta 23 juta kata menghasilkan bahwa LDA mendapatkan nilai koherens lebih baik yakni 0.59179 dengan jumlah 20 topik,

sedangkan LSA mendapatkan 10 topik dengan nilai koherens sebesar 0.577302 (Mohammed and Al-Augby, 2020). Pada penelitian *topic modeling* berbasis twitter *hashtag* menghasilkan algoritma LDA lebih baik dengan nilai koherensi sebesar 0.6047, sedangkan pada algoritma LSA mendapatkan nilai koherensi sebesar 0.4744 (Alash and Al-Sultany, 2020).

Berdasarkan dari paparan latar belakang yang telah dijelaskan, dilakukannya analisis pemodelan topik pada opini masyarakat pengguna twitter pada PT Kereta Api Indonesia (Persero) menggunakan metode *Latent Dirichlet Allocation* (LDA).

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, maka dapat dirumuskan permasalahan sebagai berikut:

1. Bagaimana hasil akurasi pemodelan topik dengan pembobotan kata *term frequency* dan *term frequency inverse document frequency*
2. Bagaimana hasil evaluasi pemodelan topik opini masyarakat terhadap PT Kereta Api Indonesia (Persero)?
3. Bagaimana hasil pengelompokan topik opini masyarakat terhadap PT Kereta Api Indonesia (Persero) menggunakan metode LDA?

1.3. Tujuan Penelitian

Tujuan yang ingin dicapai pada penelitian ini diantara lain:

1. Mengetahui metode pembobotan kata terbaik dalam pemodelan topik menggunakan metode LDA
2. Mengetahui hasil evaluasi dalam pemodelan topik menggunakan metode

LDA

3. Mengetahui topik apa saja mengenai opini masyarakat terhadap oleh pengguna jasa pelayanan PT Kereta Api Indonesia (Persero)

1.4. Manfaat Penelitian

Adapun manfaat dari penelitian ini bagi masyarakat yaitu:

1. Secara Teoritis

   Secara teoritis penelitian ini diharapkan dapat menambah pengetahuan dan wawasan mengenai pemodelan topik menggunakan *latent dirichlet allocation*.

2. Secara Praktis

   (a) Bagi Universitas

   Bagi Universitas, penelitian ini diharapkan dapat memperbanyak koleksi bahan bacaan dan referensi belajar yang bermanfaat bagi mahasiswa dan mahasiswi Prodi Matematika UINSA Surabaya.

   (b) Bagi Penulis

   Bagi penulis, penelitian ini dapat mengaplikasikan ilmu yang telah dipelajari.

   (c) Bagi Pembaca

   Bagi pembac, penelitian ini diharapkan dapat digunakan sebagai referensi keilmuan dan menambah wawasan mengenai pemodelan topik menggunakan *latent dirichlet allocation*.

1.5. Batasan Masalah

Batasan masalah dalam penelitian ini yaitu:

1. Data yang digunakan berupa opini masyarakat yang diperoleh melalui twitter pada periode 01 Januari 2022 hingga 31 Desember 2022
2. Data yang digunakan berupa *tweet* dengan kata pencarian ”@KAI121”
3. Pemodelan topik menggunakan bahasa indonesia
4. Bahasa asing, bahasa daerah dan emoji pada *tweet* diabaikan

1.6. Sistematika Penulisan

Sistematika penulisan dirancang untuk mengetahui tahapan/gambaran atau skema penelitian. Dengan kata lain, secara sederhana sistematika penulisan adalah tata cara atau kerangka dari suatu penelitian. Sistematika penulisan pada skripsi ini terbagi menjadi 5 bab, yaitu sebagai berikut.

1. BAB I

   Bab I berisi mengenai paparan pendahuluan, dimana dalam pendahuluan terbagi menjadi beberapa subbab, yaitu latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah, dan sistematika penulisan.

2. BAB II

   Bab II berisi mengenai pemaparan tinjauan pustaka, dimana dalam tinjauan pustaka berisi pemaparan teori-teori yang relevan dengan topik utama dalam penelitian ini. Dalam penelitian ini, tinjauan pustaka terbagi menjadi 11 subbab, yaitu PT Kereta Api Indonesia (Persero), Twitter, *Data Samlpling*,

Text Mining, *Text Preprocessing*, Pembobotan Kata yang terdiri dari 2 bagian yakni *Term Frequency* dan *Term Frequency Inverse Document Frequency*, Pemodelan Topik, *Latent Dirichlet Allocation*, Nilai Koherensi. *Word Cloud*, dan integrasi keislaman.

3. BAB III

   Bab III berisi mengenai pemaparan metode penelitian, dimana dalam metode penelitian meliputi jenis penelitian yang digunakan, sumber data, dan tahapan penelitian.

4. BAB IV

   Bab IV berisi mengenai pemaparan hasil dan pembahasan, yang meliputi deskripsi data, perhitungan pembobotan kata, pembentukan topik, evaluasi model uji coba, analisis model topik, dan integrasi keislaman.

5. BAB V

   Bab V berisi mengenai pemaparan penutup dari penelitian, yang meliputi kesimpulan dari hasil dan pembahasan penelitian yang telah dipaparkan, dan saran untuk pengembangan penelitian selanjutnya.

BAB II

TINJAUAN PUSTAKA

2.1. PT Kereta Api Indonesia (Persero)

PT Kereta Api Indonesia (Persero) merupakan perusahaan penyedia jasa transportasi kereta api di bawah kewenangan badan usaha milik negara (BUMN). Awal mula perkeretaapian di Indonesia sudah ada sejak penjajahan Belanda pada tahun 1864 dengan jalur pertamanya ialah Solo - Yogyakarta. Lalu pada tahun 1994 diambil oleh pemerintahan Jepang pada tahun 1942. Setelah Indonesia merdeka kuasa tersebut seutuhnya diambil oleh negara dengan nama Djawatan kereta Api Republik Indonesia (DKARI). Pada tahun 1950 nama DKARI diganti menjadi PNKA (Perusahaan Negara Kereta Api). Selanjutnya pemerintah mengganti nama PNKA menjadi Perusahaan Jawatan Kereta Api Indonesia (PJKA) pada tahun 1971. Lalu lahirlah PT Kereta Api Indonesia (Persero) pada tahun 1998.

Saat ini kementerian perhubungan (kemenhub) mengatakan pada tahun 2020 panjang rel kereta di Indonesia mencapai 6,32 juta meter, dengan mayoritas beroperasi di pulau jawa. PT Kereta Api Indonesia (Persero) saat ini memiliki tiga produk layanan yakni layanan angkutan penumpang, layanan angkutan barang, dan layanan pengusahaan aset.

2.1.1. Layanan Angkutan Penumpang

Sebagai perusahaan yang mengatur pengoperasian kereta api (KA) di Indonesia, PT Kereta Api Indonesia (Persero) telah banyak menjalankan KA penumpangnya. Untuk penggunaan angkutan penumpang KA dirancang untuk mencangkup jarak dekat, jarak menengah, hingga jarak jauh. KA tersebut dibagi menjadi beberapa kelas diantaranya kelas luxury, kelas eksekutif, kelas bisnis, dan kelas ekonomi.

2.1.2. Layanan Angkutan Barang

Selain kereta angkutan penumpang, PT Kereta Api Indonesia (Persero) juga memiliki KA angkutan barang. Angkutan KA barang terdiri angkutan bahan bakar minyak (BBM). angkutan peti kemas, angkutan batu bara, dan lain sebagainya.

2.1.3. Layanan Pengusahaan Aset

PT Kereta Api Indonesia dalam mengelola bisnisnya tidak hanya memberikan pelayanan angkutan saja, terdapat layanan pengusaha aset. layanan pengusaha aset terbagi menjadi tiga bagian yakni stasiun, non stasiun, dan iklan.

2.2. Twitter

Twitter merupakan salah satu media sosial yang sering digunakan oleh masyarakat. Pada awal kemunculan tahun 2006 dalam sekali unggahan twitter paling banyak berisi 140 karakter (Schmidt, 2014). Namun sejak tahun 2018, pengguna dapat membuat satu postingan twitter berisi hingga 280 karakter. Dengan batasan tersebut membuat twitter sebagai sebuah layanan *micro-blogging*. Selain melihat tweet orang yang diikuti terdapat beberapa fitur pada twitter yang

membedakan dengan media sosial yang lainnya, diantaranya yaitu (Kwak et al., 2010):

1. Menandai pengguna lain dalam sebuah unggahan (tweet). Dengan menggunakan simbol @ lalu dilanjutkan nama pengguna seperti @namapengguna.
2. Hashtag (#) merupakan kata kunci atau topik dalam tweet. Pengguna dapat membuat hashtag apa saja dan dapat mencari tweet dengan topik tertentu menggunakan # lalu dilanjutkan dengan kata yang mewakili.
3. Retweet (RT) merupakan langkah untuk memposting ulang atau menyebarkan tweet pengguna lain ke semua pengikut anda.

Seiring berjalannya waktu twitter tidak hanya digunakan sebagai postingan keseharian belaka, banyak sekali penelitian-penelitian yang memanfaatkan data data tweet yang ada pada twitter sebagai bahan penelitian (Bosnjak et al., 2012). Banyak berbagai cara yang digunakan untuk mendapatkan data tersebut salah satunya dengan menggunakan *TwitterAPI*.

2.2.1. Twitter *Application Programming Interface*

Twitter *Application Programming Interface* (API) merupakan salah satu akses untuk dapat mengambil data di twitter oleh pengguna umum (Brzustewicz and Singh, 2021). Dengan menggunakan twitter API akan mempermudah dalam menganalisis data twitter dengan memasukkan seluruh data twitter yang diinginkan lalu memasukannya ke dalam file berbentuk json (Kearney, 2019).

2.3. *Data Sampling*

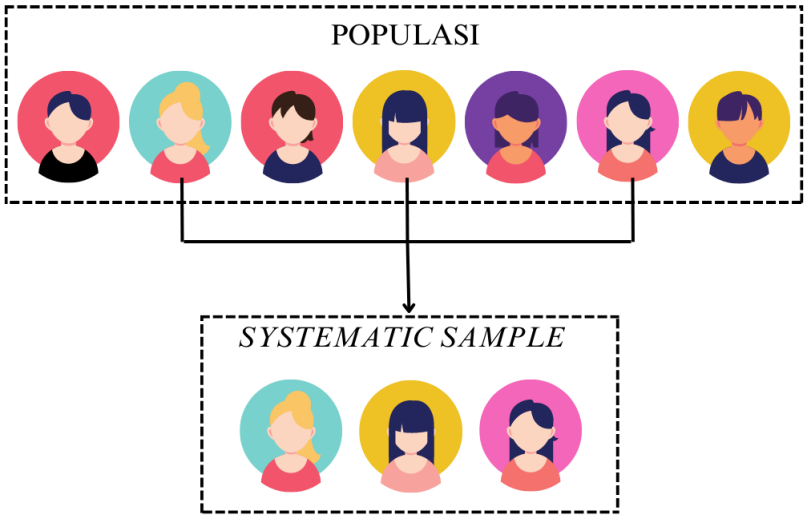
Sentimen data twitter merupakan dataset dengan jumlah yang cukup banyak dengan waktu yang singkat. *Data sampling* dapat mempermudah dalam perhitungan komputasi dengan mengurangi jumlah dataset tanpa merubah struktur dari data set aslinya. *Data sampling* adalah proses pemilihan unit dari populasi yang diminati. Dengan kata lain, *data sampling* adalah proses pemilihan subset (sampel) dari kumpulan data asli dalam jumlah besar untuk dianalisis dalam pembuatan model (Albattah, 2016).

Dalam proses *data sampling*, untuk menguatkan hasil sampel tersebut perlu diperhatikan faktor faktor yang mendukung. Diantaranya adalah metodologi pengambilan sampel, ukuran sampel, dan tingkat respon (Acharya et al., 2013). Dalam pengaplikasiannya data sampling dibagi menjadi dua kelas yakni *non-probability sample*, dan probability sample. *Non-probability sample* adalah pengambilan sampel dimana probabilitas subjek yang dipilih tidak diketahui dan menghasilkan bias seleksi dalam penelitian. Contoh *non-probability sample* adalah *purposive sampling*, *quota sampling*, *snowball-sampling* dll (Acharya et al., 2013).

Probability sample merupakan salah satu metode dalam mengambil sampel data dengan standar yang tinggi. Dalam *probability sample* setiap individu dalam populasi memiliki peluang yang sama untuk dipilih dalam penelitian. Contoh metode metode dalam *probability sample* diantaranya *simple random sampling*, *cluster sampling*, *stratified random sampling*, *systematic sampling* dan lain sebagainya (Acharya et al., 2013).

2.3.1. *Systematic Sampling*

Systematic sampling pertama dipelajari oleh Madows pada tahun 1994. Metode *Systematic Sampling* mengambil stratifikasi yang jelas atau tersembunyi dalam populasi sehingga lebih tepat daripada pengambilan sampel secara acak (Wolter, 1984). Dalam menentukan sampel, pemilihan subjek pertama akan dipilih secara acak lalu selanjutnya dilakuan secara periodik hingga data terakhir (Acharya et al., 2013). Prinsip *systematic sampling* dapat dilihat pada Gambar 2.1



Gambar 2.1 Konsep *Systematic Sampling*

Systematic sampling biasanya disebut juga sebagai sampel acak sederhana semu karena memiliki sifat yang mirip dengan sampel acak sederhana. Pemilihan subjek pertama di misalkan pada data ke-n dan selanjutnya pemilihan data akan dipilih setiap data berkelipatan ke-n (Henry, 2016). Seperti contohnya pada Gambar 2.1, pada gambar tersebut *systematic sampling* mengambil sampel data setiap kelipatan 2 dari data populasi.

2.4. *Text Mining*

Teks merupakan data tidak terstruktur yang terdiri dari beberapa kata. *Text Mining* merupakan suatu proses yang berfungsi untuk mencari informasi dari data yang berupa teks. Tujuan utama dari *text mining* adalah untuk menemukan dan menganalisis pola data, termasuk trend dan outlier dalam data teks, bahkan gagasan yang relevan maupun tidak relevan dalam jumlah besar (Jo, 2019). Text mining juga mampu mengkategorikan teks (*text categorization*) dan pengelompokan teks (*text clustering*).

Dalam prosesnya *text mining* menngembangkan dan menggunakan beberapa teknik dari bidang lain, seperti matematik, statistik, linguistik, *Natural Language Processing* (NLP), kecerdasan buatan, dan lain sebagainya. Adapun fitur dari text mining diantaranya yaitu text processing, analisis sentimen, *text analytics*, klasifikasi/kategorisasi, *knowledge discovery*, dan lain sebagainya (Kaur and Chopra). *Text mining* mempunyai beberapa tahap dalam pengolahan datanya diantaranya *text preprocessing* yaitu proses awal terhadap text, transformasi teks (*text transformation*), penemuan pola (*pattern discovery*), dan evaluasi (*evaluate*) (Kumar and Bhatia, 2013).

2.5. *Text Preprocessing*

Text preprocessing merupakan tahap pertama yang dilakukan dalam menganalisis teks. Tahap ini sangat penting dilakukan karena untuk membuang kata-kata yang tidak dibutuhkan di dalam dokumen. Pada tahap ini sangat mempengaruhi hasil setelah dilakukan pengolahan data teks. Tahap *preprocessing* bertujuan untuk mengubah data yang tidak terstruktur atau semi terstruktur menjadi model ruang terstruktur. Jika pada proses ini dilakukan secara baik maka

hasil yang diperoleh juga akan baik, begitu pula sebaliknya. Proses *preprocessing* terdiri dari beberapa proses yaitu sebagai berikut:

2.5.1. *Remove Punctuation*

Langkah pertama pada *preprocessing* adalah *remove punctuation*. Pada tahap ini menghilangkan seluruh karakter yang tidak digunakan dan tidak memiliki nilai analitik (El Rahman et al., 2019). (&,,%, dan sebagainya), angka, hashtag, url, nama pengguna twitter, RT, dan *emoticon*. Karakter-karakter tersebut biasanya dikenal sebagai noise. Noise adalah suatu bentuk data dimana menimbulkan potensi yang akan mengganggu jalannya proses pengolahan.

2.5.2. *Case Folding*

Langkah berikutnya setelah *remove punctuation* adalah tahap *case folding*. Proses ini akan mengubah seluruh karakter huruf besar menjadi huruf kecil (*lowercase*) (Erman and Sitanggang, 2016). Tahap ini perlu dilakukan, karena pada dasarnya bahasa komputer menerima kata menggunakan huruf kecil dan huruf besar sebagai informasi yang berbeda, walaupun sebenarnya adalah kata yang sama. Sehingga tahap ini dapat menghindari terdapatnya duplikat kata yang sama.

2.5.3. *Tokenizing*

Tahap *preprocessing* berikutnya adalah *tokenizing*. *Tokenizing* bertujuan untuk memisahkan suatu teks menjadi kata (token) individual (Pratiwi, 2022) Proses ini berguna untuk mempermudah dalam penghitungan kata sampai dengan transformasi kata menjadi vektor berdimensi tinggi.

2.5.4. *Slang Word Changer*

Slang word adalah kata dan frasa yang digunakan pada kehidupan sehari-hari untuk membangun identitas sosial dengan tren atau model dalam masyarakat pada umumnya, atau biasa disebut kata gaul (Trimastuti, 2017). Pada tahap ini akan mengubah bentuk kata gaul menjadi kata sesuai pada kamus besar bahasa indonesia (KBBI).

2.5.5. *Correcting Word*

Correcting word adalah tahap membenarkan kata kata yang memiliki kesalahan penulisan. Kata yang tidak masuk ke dalam kata gaul dan KBBI dianggap sebagai kata dengan kesalahan penulisan (Setiabudi et al., 2021). Kata tersebut akan dihitung per huruf sesuai dengan dalam kata pada kamus. Kesalahan penulisan dapat menyebabkan perluasan kata yang tidak perlu sehingga dapat mengubah hasil pengolahan.

2.5.6. *Stemming*

Stemming merupakan proses menormalkan token kata menjadi satu bentuk kata dasar atau bisa disebut langkah mencari akar kata (*root word*) dari setiap kata. Setiap imbuhan kata akan digunakan sebagai kata dasar. Proses ini sangat penting dilakukan, karena bahasa Indonesia memiliki struktur yang kompleks dimana dalam sebuah kata memiliki berbagai macam imbuhan (Rohman and Asror, 2019).

2.5.7. *Stopwords*

Menurut definisi *stopword* merupakan kata-kata yang tidak memiliki makna (Saif et al., 2014). Sehingga *stopwords* dalam tahap *preprocessing* adalah

proses menghilangkan kata yang tidak digunakan pada dokumen dengan melakukan pengecekkan kata kata hasil tersebut termasuk dalam kata yang tidak penting (*stoplist*) atau tidak. Kata kata tersebut biasanya terdiri dari kata depan, kata hubung, dan kata yang tidak penting menurut peneliti. Kata yang termasuk dalam stoplist akan dihapus sehingga yang tersisa hanya kata kata yang dianggap penting (*keywords*).

2.6. Pembobotan Kata *(Term Weighting)*

Pembobotan kata perlu dilakukan guna untuk memberikan bobot atau nilai dari kata yang terkandung di dalam dokumen. Dengan kata lain pembobotan kata adalah proses mengubah data teks menjadi data numerik. Dalam menetapkan bobot per kata terdapat dua skema yakni pembobotan dengan tidak memperhatikan urutan kata dalam suatu dokumen (*Unsupervised term weightin*g), dan pembobotan dengan memperhatikan urutan kata dalam suatu dokumen (*Supervised term weightin*g) (Onan and Tocoglu, 2021).

Pada *unsupervised term weighting* diasumsikan bahwa jika diberi M dokumen di dalam korpus, TF (*term frequency*) menunjukkan jumlah kata (*term*) di dalam dokumen. Sementara, DF (*Document Frequency*) merupakan jumlah dokumen dimana kata (*term*) tersebut itu muncul (Samant et al., 2019). Untuk mencari bobot menggunakan TF dapat dilihat pada persamaan 2.1.

$$w_{w,v} = tf_{w,v} \tag{2.1}$$

Dimana :

$w_{w,v}$: Bobot kata v pada dokumen w

$tf_{m,v}$: Jumlah kemunculan kata v pada dokumen w

Selain TF, metode *unsupervised term weighting* lainnya yakni *Term Frequency-Inverse Document Frequency* (TF-IDF). TF-IDF menggabungkan dua konsep yakni frekuensi kemunculan kata dalam suatu dokumen dan inverse frekuensi dokumen yang mengandung kata tersebut. Skema pembobotan TF-IDF dapat dihitung menggunakan persamaan 2.2

$$w_{w,v} = tf_{w,v} \times \log\left(\frac{M}{df_j}\right) \tag{2.2}$$

Dimana:

$w_{w,v}$: Bobot kata v pada dokumen w

$tf_{w,v}$: Jumlah kemunculan kata v pada dokumen w

df_v : Jumlah dokumen yang mengandung kata v

M : Jumlah dokumen

2.7. Pemodelan Topik

Pemodelan topik adalah salah satu teknik dengan *unsupervised learning* yang dapat mengidentifikasi tema dari kumpulan beberapa dokumen tertentu dan menemukan distribusi topik dari setiap dokumen (Sarioglu et al., 2013). Ide dasar pemodelan topik adalah bahwa dokumen terdiri dari beberapa topik, dimana topik sendiri merupakan distribusi probabilitas dari kata kata. Kumpulan dokumen yang mempunyai distribusi probabilitas topik juga memiliki setiap kata yang mewakili satu topik. Sehingga adanya distribusi probabilitas topik pada setiap dokumen, dapat mengetahui jumlah topik yang terlibat dalam suatu dokumen. Dari hal

tersebutlah dapat diketahui topik mana yang yang dibicarakan dokumen tersebut (Steyvers and Griffiths, 2010).

2.8. *Latent Dirichlet Allocation* (LDA)

Latent Dirichlet Allocation (LDA) adalah salah satu metode pemodelan topik yang populer saat ini. Ide dasar LDA yakni bahwa adanya topik acak tersembunyi dalam sebuah dokumen, dan topik tersebut tersusun atas kumpulan kata - kata (Blei, 2012). Secara formal dapat didefinisikan sebagai berikut :

1. Kata merupakan bentuk terkecil dari data diskrit yang telah diberi indeks $(1, 2, ...N)$. Dimana N Merupakan jumlah token kata dalam satu korpus.
2. Sebuah dokumen merupakan barisan dari kata kata V yang dinotasikan sebagai $w = (v_1, v_2, ..., v_V)$.
3. Sebuah korpus adalah koleksi dari M dokumen dinotasikan dengan $D = (w_1, w_2, ..., w_M)$

LDA merupakan model probabilitas generatif dari sekumpulan data diskrit seperti kumpulan teks (Blei et al., 2003). Dalam pemodelan generatif, data yang dihasilkan dari proses generatif termasuk variabel tersembunyi (laten). Proses generatif ini diartikan sebagai distribusi probabilitas gabungan atas variabel acak yang diamati dan variabel laten. Analisis data menggunakan distribusi bersama untuk menghitung distribusi bersyarat dari struktur topik (variabel bersyarat) dengan variabel teramat. Distribusi bersyarat dikenal juga sebagai distribusi posterior.

Sebagai model grafis probabilistik, dalam algoritma LDA memiliki tiga tingkat di dalamnya, yakni kata, topik, dan dokumen. Dalam menentukan

distribusi topik dalam dokumen menggunakan parameter (α). Sedangkan untuk menentukan distribusi kata pada topik menggunakan parameter (β). Distribusi topik dalam dokumen (α) mengakibatkan munculnya variabel Θ sebagai kumpulan topik. Proses generatif tersebut diasumsikan sebagai setiap D korpus yang terdiri dari dokumen sebanyak M dengan langkah-langkah berikut (Jelodar et al., 2018):

1. Untuk setiap dokumen $w(w \in \{1, 2, 3, \cdots, M\})$, pilih distribusi topik multinomial (Θ_w) dari distribusi dirichlet dengan parameter α dirumuskan sebagai berikut:

$$\Theta_w \sim Dirichlet(\alpha)$$

2. Untuk setiap topik $k(k \in \{1, 2, 3, \cdots, K\})$, pilih distribusi kata multinomial Φ_k dari distribusi dirichlet menggunakan parameter β dirumuskan sebagai berikut:

$$\Phi_k \sim Dirichlet(\beta)$$

3. Untuk setiap N kata dalam dokumen $w, (v \in \{1, 2, 3, \cdots, N_w\})$

   (a) Pilih topik z_v dari Θ_n atau dapat dirumuskan sebagai berikut:

$$z_v \sim Multinomial(\Theta_w)$$

   (b) pilih kata dari w_v dari $\Phi_{k,v}$ atau dapat dirumuskan sebagai berikut:

$$w_v \sim Multinomial(\Phi_{k,v})$$

Variabel Θ adalah distribusi topik terhadap dokumen. Variabel z adalah gambaran topik dari kata tertentu dalam sebuah dokumen. Dan variabel w merupakan representasi dari kata yang berkaitan dengan topik yang ada dalam dokumen.

LDA adalah model yang secara acak dibangkitkan oleh data observasi (variabel teramati) yang didalamnya terdapat variabel laten (variabel tersembunyi). Data observasi merupakan kumpulan kata dari beberapa dokumen (W). Kumpulan kata dalam dokumen di dalam korpus dibentuk secara matriks yang didapatkan dari hasil pembobotan kata.

$$W = \begin{bmatrix} w_{1,1} & \cdots & w_{1,V} \\ \vdots & \ddots & \vdots \\ w_{M,1} & \cdots & w_{M,V} \end{bmatrix}_{M \times V} \tag{2.3}$$

dimana,

M : Jumlah dokumen dalam korpus

V : Jumlah kata dalam korpus

Sedangkan variabel laten ialah topik yang telah ditetapkan oleh sekumpulan kata di dalamnya (Z). Penentuan awal topik dilakukan secara acak, dengan didefinisikan sebagai berikut:

$$Z_k = \begin{bmatrix} z_{1,1} & \cdots & z_{1,V} \\ \vdots & \ddots & \vdots \\ z_{M,1} & \cdots & z_{M,V} \end{bmatrix}_{M \times V} \tag{2.4}$$

Lalu variabel laten tersebut diperbarui menggunakan pembangkitan model LDA dilakukan oleh distribusi probabilitas masing masing parameter, sehingga

membentuk sebuah distribusi posterior (distribusi bersama). Selanjutnya menghidupkan distribusi probabilitas topik terhadap dokumen yang memiliki distribusi dirichlet dengan parameter α dapat dilihat pada persamaan 2.5.

$$p(\Theta_w|\alpha) = \frac{\Gamma\left(\Sigma_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\prod_{k=1}^{k}\Theta_k^{a_{k-1}} \tag{2.5}$$

dimana,

α : Parameter Dirichlet Dokumen

Θ_k : Probabilitas topik k

k : Topik ke-k

K : Jumlah topik

Berikutnya menggunakan distribusi probabilitas kata kata terhadap topik yang telah ditentukan pada masing masing dokumen menggunakan peluang bersyarat untuk Φ dan menggunakan parameter β. Dapat dilihat pada persamaan 2.6.

$$p(\Phi_k|\beta) = \frac{\Gamma\left(\Sigma_{v=1}^{N}\beta_k\right)}{\prod_{v=1}^{N}\Gamma\left(\beta_{k,v}\right)}\prod_{v=1}^{N}\Phi_{k,v}^{\beta_{k,v-1}} \tag{2.6}$$

dimana,

$\Phi_{k,v}$: Probabilitas kata ke-v pada topik ke-k

β : Parameter Dirichlet kata

k : Topik ke-k

K : Jumlah topik

V : Jumlah kata

Lalu membangkitkan distribusi probabilitas topik yang telah ditentukan tiap dokumen berdasarkan pada kata yang muncul di dalamnya, menggunakan distribusi multinomial. Distribusi multinomial dapat dilihat pada persamaan 2.7.

$$p(z_{w,v}|\Theta_w) = \prod_{w=1}^{M}\prod_{k=1}^{K}\Theta_{w,k}^{n_{w,k}} \tag{2.7}$$

dimana,

$n_{w,k}$: Jumlah kemunculan topik k yang ditetapkan oleh kata kata dari dokumen w.

Selanjutnya menghidupkan distribusi probabilitas untuk kata kata dalam korpus terhadap topik yang telah dipilih dengan menggunakan persamaan 2.8.

$$p(w_{w,v}|z_{w,v}, \Phi_{w,k}) = \prod_{k=1}^{K}\prod_{v=1}^{N}\Phi_{k,v}^{n_{k,v}} \tag{2.8}$$

dimana,

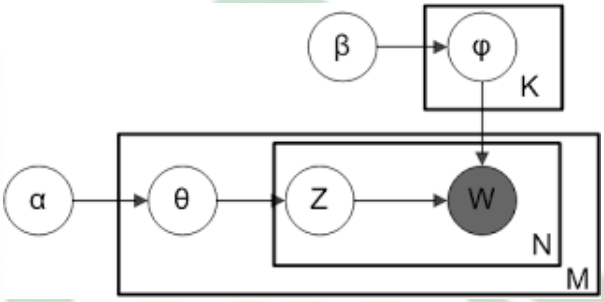
$n_{k,v}$: Jumlah kemunculan topik ke-k yang ditentukan dari kata-kata pada korpus.

Setelah menghitung seluruh distribusi variabel tersembunyi,selanjutnya untuk proses iterasi menggunakan distribusi posterior. Proses generatif distribusi posterior LDA dapat dilihat pada persamaan 2.9. Masalah utama dalam pemodelan topik adalah inferensi distribusi posterior. Dalam perhitungan komputasi, menghitung inferensi dari LDA dari struktur topik di mana pembilangnya adalah distribusi posterior dari semua variabel acak dinyatakan pada persamaan 2.10 (Reisenbichler and Reutterer, 2019).

$$p(\Theta, \Phi, z, w|\alpha\beta) = \prod_{w=1}^{M} p(\Theta|\alpha) \prod_{k=1}^{K} p(\Phi|\beta) \prod_{v=1}^{N} p(z|\Theta)p(w|\Phi, z) \quad (2.9)$$

$$p(\Theta, \Phi, z|w, \alpha, \beta) = \frac{p(\Theta, \Phi, z, w|\alpha\beta)}{p(w|\alpha, \beta)} \quad (2.10)$$

Jika divisualisasikan algoritma LDA dapat dilihat pada Gambar 2.2



Gambar 2.2 Grafik Model Algoritma LDA

Seperti yang sudah dijelaskan sebelumnya, dari Gambar 2.2 terdapat beberapa inisial. α mempresentasikan sebagai parameter dirichlet dari topik dalam dokumen. Selanjutnya inisial Θ merupakan distribusi multinomial dari probabilitas topik setiap dokumen. Sementara Z adalah presentasi dari probabilitas topik terhadap topik yang ditentukan. Dan β merupakan parameter dirichlet dari kata pada topik. Φ distribusi multinomial dari probabilitas kata kata terhadap topik yang telah ditentukan. Dan W merupakan probabilitas untuk kata di dalam korpus terhadap topik.

Pada gambar 2.2 terdapat bentuk persegi dengan keterangan K merupakan perulangan sejumlah banyaknya topik. Sementara bentuk persegi dengan keterangan N merupakan perulangan sejumlah banyaknya jumlah kata. Dan persegi dengan keterangan M merupakan perulangan sejumlah banyaknya dokumen

2.8.1. *Collapsed Gibbs Sampling*

Pada pengaplikasian model LDA, perlu dilakukannya estimasi parameter. Karena LDA tidak dapat menemukan variabel laten secara langsung. Estimasi parameter tersebut dapat dilakukan menggunakan algoritma *gibbs sampling*. *Gibbs Sampling* merupakan algoritma salah satu bagian dari kerangka kerja *Markov Chain Monte Carlo* (MCMC).

Algoritma MCMC bertujuan untuk membangun rantai Markov yang memiliki distribusi posterior target sebagai distribusi stasioner. Dengan kala lain, setelah beberapa iterasi melalui rantai tersebut, pengambilan sampel dari distribusi harus mendekati pengambilan sampel dari distribusi posterior yang diinginkan (Darling, 2011). Sebagai contoh, dalam mengambil sampel x dari distribusi bersama $p(x) = p(x_1, x_2, ..., x_m)$, dimana tidak ada solusi tertutup untuk $p(x)$. Namun, representasi untuk distribusi kondisional yang terkait. Dalam hal ini, dapat menggunakan metode *gibbs sampling* dengan langkah-langkah berikut (Maindonald, 2007):

1. Inisialisasi secara acak setiap x_1.
2. Untuk setiap iterasi $t - 1, 2, ...T$:
   (a) $x_1^{t+1}\ p(x_1|x_2^{(t)}, x_3^{(t)}, ..., x_m^{(t)})$
   (b) $x_2^{t+1}\ p(x_2|x_1^{(t+1)}, x_3^{(t)}, ..., x_m^{(t)})$
   (c) $x_m^{t+1}\ p(x_m|x_1^{(t+1)}, x_2^{(t+1)}, ..., x_{m-1}^{(t+1)})$

Langkah-langkah tersebut diulang hingga sampel mencapai hasil yang konvergen dengan distribusi yang sebenarnya. *Gibbs sampling* terbukti memiliki kinerja yang cukup kuat dalam mengestimasi konvergensi. Oleh karena itu

algoritma LDA dapat diturunkan untuk setiap variabel tersembunyi. Yakni, dokumen distribusi topik terhadap dokumen (Θ) maupun distribusi kata terhadap topik (Φ) dapat dihitung hanya dengan menggunakan alokasi indeks topik z. Sehingga, dapat menggunakan algoritma yang lebih sederhana jika diintegrasikan dengan parameter multinomial dan hanya melakukan sampling pada nilai z. Pendekatan tersebut disebut dengan *collapsed gibbs sampling*. Dalam pengambilan sampel algoritma *collapsed gibbs sampling* menggunakan beberapa matriks, diantaranya yakni (Papanikolaou et al., 2017):

1. Matriks $TK_{K\times V}$ adalah representasi dari jumlah bobot kata v pada topik k diseluruh korpus.
2. Matriks $KD_{W\times V}$ adalah representasi dari jumlah kata pada dokumen w yang ditetapkan pada topik k.

Matriks tersebut berfungsi sebagai *input* sampel topik pada setiap kata dan dokumen. Setelah sampel topik tersebut ditetapkan kemudian akan dilakukan pengulangan perhitungan menggunakan distribusi posterior. Distribusi posterior LDA *collapsed gibbs sampling* merupakan perkalian antara distribusi kata terhadap topik dengan distribusi probabilitas topik terhadap dokumen yang diintegrasikan menggunakan persamaan 2.10 sehingga mendapatkan yang dirumuskan pada persamaan 2.11. Proses pengulangan perhitungan dilakukan hingga mencapai jumlah iterasi maksimal yang diinginkan. Sementara itu, untuk menghitung distribusi kata pada setiap topik menggunakan persamaan 2.12. Sedangkan untuk menghitung probabilitas topik terhadap dokumen menggunakan persamaan 2.13

$$p(z_i = k|w_i = v, d, \alpha, \beta) \propto \frac{TK_{k,v} + \beta}{\sum_{v=1}^{N} TK_{k,v} + \beta N} \frac{KD_{w,k} + \alpha}{\sum_{k=1}^{K} DK_{w,k} + \alpha K} \quad (2.11)$$

$$\Phi^{vk} = \frac{TK_{k,v} + \beta}{\sum_{v=1}^{N} TK_{k,v} + \beta N} \quad (2.12)$$

$$\Theta^{kw} = \frac{KD_{w,k} + \alpha}{\sum_{k=1}^{K} DK_{w,k} + \alpha K} \quad (2.13)$$

dimana,

z_i=k : Pembaruan topik untuk kata v topik k

Φ^{vk} : Probabilitas kata v pada topik k

Θ^{kw} : probabilitas topik k pada dokumen w

w_i=v : Kata yang diproses

d : Dokumen yang di proses

α : Parameter dirichlet atas distribusi topik

β : Parameter dirichlet atas distribusi kata

$TK_{k,v}$: Bobot kata v yang ditempatkan pada topik k tanpa termasuk kata v

$KD_{d,k}$: Jumlah kemunculan topik k ditempatkan ke dalam token kata pada dokumen d tanpa termasuk kata v

N : Jumlah kata

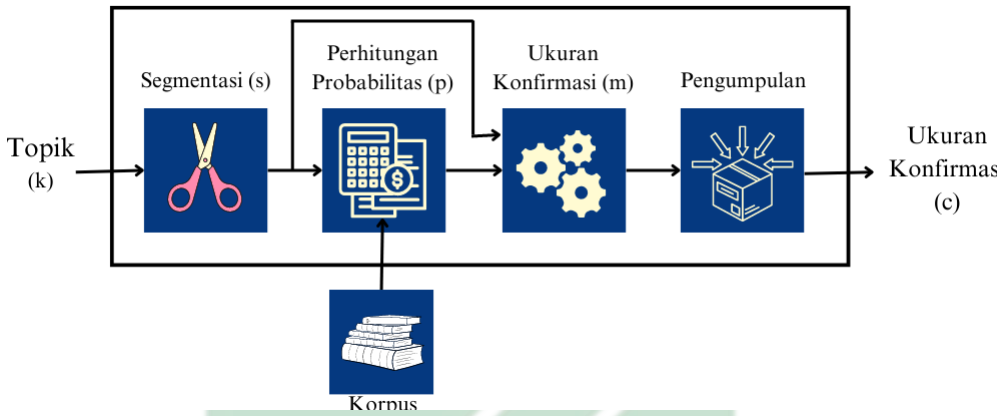
K : Jumlah topik

2.9. Nilai Koherensi

Dalam mengukur baik atau buruknya suatu model dalam pemodelan topik diperlukan evaluasi. Evaluasi tersebut dapat mengukur bagaimana pendekatan kata kata yang terdapat di dalam suatu topik dengan menggunakan independensi

statistik. Metode tersebut disebut dengan nilai koherensi (Kherwa and Bansal, 2020). Sebagai contoh jika suatu dokumen membahas fashion, terdapat topik dengan kata kata (baju, topi, dan sepatu) akan terlihat baik. Dengan kata lain, kata kata tersebut saling berkorelasi satu dengan yang lainnya.

Nilai koherensi diartikan sebagai rata-rata atau median kesamaan kata berpasangan yang dibentuk oleh kata-kata teratas dari topik tertentu (Rosner et al., 2014). Secara garis besar, evaluasi ini terdiri dari beberapa bagian. Ilustrasi struktur tahap nilai koherensi dapat dilihat pada Gambar 2.3



Gambar 2.3 Struktur Tahapan Nilai Koherensi

Dapat dilihat pada 2.3, untuk mendapatkan nilai koherensi perlu dilakukannya 4 tahapan. Pada dasarnya nilai koherensi memiliki algoritma yang berbeda beda. Salah satu algoritma terbaik untuk mencari nilai koherensi yang paling mendekati dengan evaluasi topik menggunakan pikiran manusia adalah koherensi C_v (Röder et al., 2015). Sebagai contoh dalam suatu korpus memiliki 4 dokumen sebagai berikut :

dokumen 1: ‘putri’, ‘beli’, ’dua’, ‘lampu’

dokumen 2: ‘putri’, ’beli’, ‘mawar’

dokumen 3: ‘mawar’, ‘warna’, ‘merah’

dokumen 4: ‘violet’, ‘warna’, ‘biru’ ‘muda’

Setelah melalui perhitungan model LDA ditemukan 3 topik dengan 2 kata utama, dapat dilihat sebagai berikut:

Topik 1: (0.3*‘mawar’ + 0.2*’merah’)

Topik 2: (0.1*‘putri’ + 0.2* ‘mawar’)

Topik 3: (0.2*‘violet’ + 0.1*‘biru’)

Untuk langkah langkah dalam mendapatkan nilai koherensi C_v adalah sebagai berikut.

2.9.1. Segmentasi

Segmentasi merupakan proses pembentukan pasangan subset kata di dalam topik. Jika suatu topik didefinisikan sebagai berikut:

$$w = \{v_1, v_2, ..., v_N\}$$

Dengan n sebagai kata jumlah kata tertinggi dari topik, penerapan segmentasi S pada nilai koherensi C_v menghasilkan sekumpulan pasangan subset dari W. Setiap kata dipasangkan dengan setiap kata lainnya. Segmentasi tersebut disebut secara satu per satu dan didefinisikan sebagai berikut:

$$S = \{(W', W^*)|W' = \{v_i\}; W^* = \{v_j\}; v_i, v_j \in v; i \neq j\} \tag{2.14}$$

dimana,

S : Segmentasi

W' : Kata pertama

W^*: Kata berikutnya setelah kata pertama

Hasil segmentasi pada 4 dokumen contoh dan 3 topik menggunakan persamaan 2.14 adalah sebagai berikut:

S_1 = (‘mawar’,’merah’) S_2 = (‘putri’, ‘mawar’) S_3 = (‘violet’ ‘biru’)

2.9.2. Perhitungan Probabilitas

Pada tahap ini menentukan cara asal perhitungan probabilitas dari sumber data yang mendasari. Metrik koherensi menggunakan probabilitas yang diambil dari korpus tekstual. Cara perhitungan yang digunakan pada nilai koherensi C_v adalah P_{sw}. Merupakan singkatan dari *probability sliding window*, cari ini menentukan jumlah kata menggunakan *sliding window* yang bergerak di atas dokumen satu langkah pada setiap token kata.

Pada umumnya C_v menggunakan $P_{sw_{110}}$, 110 merupakan ukuran dari *sliding window tersebut*. Sebagai contoh, akan digunakan P_{sw_3} menggunakan data dan topik yang telah tersedia, sehingga didapatkan hasil sebagai berikut

sw_3 = (‘putri’, ‘beli’, ’dua’), (‘beli’, ’dua’, ‘lampu’), (‘putri’, beli, ‘mawar’), (‘mawar’, ‘warna’, ‘merah’), (‘violet’, ‘warna’, ‘biru’), (‘warna’, ‘biru’, ‘muda’)
Sehingga ditemukan nilai probabilitas katanya sebagai berikut :

P(’mawar’) = $\frac{2}{6}$, P(’merah’) = $\frac{1}{6}$, P(’putri’) = $\frac{2}{6}$, P(’violet’) = $\frac{1}{6}$,
P(’biru’) = $\frac{1}{6}$, P(’mawar,merah’) = $\frac{1}{6}$, P(’putri, mawar’) = $\frac{1}{6}$,
P(’violet, biru’) = $\frac{1}{6}$

2.9.3. Ukuran Konfirmasi

Ukuran konfirmasi merupakan tahapan inti dari mencari nilai koherensi. Tahap ini menghitung seberapa baik pasangan subset dari persamaan 2.14. Nilai

koherensi pada dasarnya berasal dari vektor yang terdiri dari beberapa kata di dalam w. Korelasi dari kata kata penyusun w tersebut yang paling mendekati pikiran manusia disebut sebagai NPMI (*Normalized Point Mutual Information*) (Aletras and Stevenson, 2013). NPMI didefinisikan pada persamaan 2.15

$$NPMI = \frac{\log \frac{P(W',W^*)+\epsilon}{P(W^*)P(W')}}{-\log P(W',W^*)+\epsilon} \tag{2.15}$$

dimana,

$P(W')$: Probabilitas kemunculan kata W'

$P(W^*)$: Probabilitas kemunculan kata W^*

$P(W',W^*)$: Probabilitas kemunculan kata W' dan W^* secara bersamaan

Perhitungan konfirmasi tersebut tidak dapat secara langsung mengasumsikan kata kata yang ada dalam topik, Sehingga diperlukan vektor sebagai berikut:

$$\vec{w}_{i,k} = NPMI(w_{j,k}^T, w_{i,k}^T), \forall i \in \{1,2,...,N\} \tag{2.16}$$

$$|\vec{w}_{i,k}| = \sqrt{{v_1}^2 + {v_2}^2} \tag{2.17}$$

Kemudian langkah terakhir pada proses ini adalah sebagai berikut :

$$S_{\cos}(\vec{w}_{n,k}, \vec{w}_{n+1,k}) = \frac{\vec{w}_{n,k}.\vec{w}_{n+1,k}}{|\vec{w}_{n,k}| \times |\vec{w}_{n+1,k}|} \tag{2.18}$$

Sebagai contoh yang telah terdapat nilai probabilitas tiap kata, langkah berikutnya mencari nilai NPMI menggunakan persamaan 2.15, mendapatkan hasil

sebagai berikut :

$$NPMI_{\text{'mawar','merah'}} = \frac{\log \frac{\frac{1}{6}}{\frac{2}{6}\times\frac{1}{6}}}{-\log \frac{1}{6}} = \frac{0.301}{0.778} = 0.387$$

$$NPMI_{\text{'mawar','merah'}} = \frac{\log \frac{\frac{1}{6}}{\frac{2}{6}\times\frac{2}{6}}}{-\log \frac{1}{6}} = \frac{0.176}{0.778} = 0.226$$

$$NPMI_{\text{'mawar','merah'}} = \frac{\log \frac{\frac{1}{6}}{\frac{1}{6}\times\frac{1}{6}}}{-\log \frac{1}{6}} = \frac{0.778}{0.778} = 1$$

Dari nilai NPMI yang telah didapatkan, sehingga dapat dihasilkan nilai dari persamaaan 2.16 sebagai berikut:

$$\vec{w}_{1,1} = (1, 0.387), \vec{w}_{2,1} = (0.387, 1)$$

$$\vec{w}_{1,2} = (1, 0.226), \vec{w}_{2,2} = (0.226, 1)$$

$$\vec{w}_{1,3} = (1, 1), \vec{w}_{2,3} = (1, 1)$$

Menggunakan konsep dasar vektor didapatkan nilai $|w|$ pada persamaan 2.17 adalah sebagai berikut:

$$|\vec{w}_{1,1}|, |\vec{w}_{2,1}| = \sqrt{1 + 0.387} = 1.072$$

$$|\vec{w}_{1,2}|, |\vec{w}_{2,2}| = \sqrt{1 + 0.226} = 1.025$$

$$|\vec{w}_{1,3}|, |\vec{w}_{2,3}| = \sqrt{1 + 1} = 1.414$$

Terakhir menghitung evaluasi tiap topik menggunakan persamaan 2.18, mendapat

nilai sebagai berikut:

$$S_{\cos}(\vec{w}_{1,1}, \vec{w}_{2,1}) = \frac{(1, 0.387)(0.387, 1)}{1.072 \times 1.072} = \frac{0.774}{1.149} = 0.674$$
$$S_{\cos}(\vec{w}_{1,2}, \vec{w}_{2,2}) = \frac{(1, 0.226)(0.226, 1)}{1.025 \times 1.025} = \frac{0.452}{1.050} = 0.430$$
$$S_{\cos}(\vec{w}_{1,3}, \vec{w}_{2,3}) = \frac{(1, 1)(1, 1)}{1.414 \times 1.414} = \frac{2}{1.999} = 1$$

2.9.4. Pengumpulan

Pada tahap ini menggabungkan seluruh nilai yang telah terkonfirmasi. Penggabungan ini menjadi nilai akhir dari nilai koherensi dengan algoritma C_v. Nilai yang sudah dikumpulkan, selanjutnya menghitung rata-rata sebagai nilai akhir. Dapat didefinisikan pada persamaan 2.19

$$C_v = \frac{\sum_{k=1}^{K}(\vec{w}_{n,k}, \vec{w}_k^*)}{K} \tag{2.19}$$

Sehingga nilai koherensi C_v pada contoh yang telah dibuat menggunakan persamaan 2.19, dapat dilihat sebagai berikut:

$$C_v = \frac{0.674 + 0.430 + 1}{3} = 0.701$$

2.10. *Word Cloud*

Untuk mempermudah data yang telah diolah untuk dipahami, penting sekali menggunakan teknik visualisasi data. Visualisasi data yang cukup digemari pada data teks salah satunya ialah *word cloud*. *Word cloud* merupakan salah satu bentuk analisis dengan metode *text mining* yang merepresentasikan kata yang sering muncul atau kata populer. Pada dasarnya *word cloud* didasarkan pada 3 jenis algoritma, yakni (Jin, 2017):

1. kategori: Pada jenis ini, ukuran setiap kata menunjukkan jumlah sub kategori koleksi. Jenis ini umumnya digunakan dalam pemetaan geografis

2. Frekuensi: Pada jenis frekuensi, ukuran setiap kata menunjukkan jumlah kata tersebut muncul. Jenis ini merupakan algoritma *word cloud* yang paling dasar.

3. Campuran: Dalam tipe ini, merupakan gabungan dari frekuensi dan kategori. Sehingga memerlukan analisis yang logis dari data yang rumit.

Untuk menampilkan hasil kata terbanyak pada suatu topik digunakannya algoritma *world cloud* yang paling dasar yakni menggunakan frekuensi kata. Penentuan ukuran kata dalam *word cloud* dapat menggunakan persamaan sebagai berikut (Jin, 2017) :

$$H_i = \begin{cases} \frac{f_{\max}(t_i - t_{\min})}{t_{\max} - t_{\min}}, & t_i > t_{\min} \\ 1, & t_i = t_{\min} \end{cases} \tag{2.20}$$

dimana,

H_i : Nilai ukuran dari kata ke-i

$f_{\max}$: Nilai ukuran kata terbesar

t_i : Jumlah kemunculan kata ke-i

$t_{\max}$: Jumlah kata terbanyak

$t_{\min}$: Jumlah kata paling sedikit

Dalam kasus dengan kata yang sangat banyak, perlu dilakukannya penskalaan. Dalam normalisasi linear, bobot t_i dari kata dipetakan ke dalam skala 1 hingga f. Contoh tampilan *word cloud* dapat dilihat pada Gambar 2.4

Gambar 2.4 Tampilan *Word Cloud*

Seperti pada gambar 2.4 semakin besar ukuran tulisan kata semakin banyak kemunculan kata. Begitu pula sebaliknya, semakin kecil ukuran tulisan kata semakin sedikit kemunculan kata.

2.11. Integrasi Keislaman

Allah menciptakan bumi dengan segala nikmat yang Allah berikan didalamnya. Nikmat tersebut dapat dirasakan di penjuru dunia. Allah memberikan Nikmat kepada hamba-Nya dalam berbagai macam bentuk. Manusia dapat bersekolah, bekerja, maupun beribadah merupakan contoh nikmat yang telah Allah berikan. Manusia menjalankan perintah Allah dan menikmati nikmat yang telah Allah berikan melalui perjalanan. Salah waktu yang dapat menyebabkan diijabahnya doa oleh Allah SWT adalah doa saat dalam perjalanan, Hal tersebut tercatat di dalam hadist yang berbunyi:

حَدَّثَنَا عَبْدُ الْمَلِكِ بْنُ عَمْرٍو حَدَّثَنَا هِشَامٌ عَنْ يَحْيَى عَنْ أَبِي جَعْفَرٍ قَالَ سَمِعْتُ أَبَا هُرَيْرَةَ يَقُولُ قَالَ رَسُولُ اللَّهِ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ ثَلَاثُ دَعَوَاتٍ مُسْتَجَابَاتٌ لَا شَكَّ فِيهِنَّ دَعْوَةُ الْمَظْلُومِ وَدَعْوَةُ الْمُسَافِرِ وَدَعْوَةُ الْوَالِدِ عَلَى وَلَدِهِ

artinya: Ada tiga doa yang mustajab (dikabulkan), tak ada keraguan padanya; doa orang terdzalimi, doa musafir, dan doa seorang bapak untuk anaknya (HR. Ahmad no 10353).

Berdasarkan hadist riwayat Ahmad no 10353 tersebut dianjurkan untuk membaca doa pada waktu yang ditentukan tersebut. Rasulullah SAW memberikan amalan untuk membaca doa saat menaiki kendaraan. Doa tersebut berbunyi :

سُبْحَانَ الَّذِىْ سَخَّرَلَنَا هَذَا وَمَاكُنَّا
لَهُ مُقْرِنِيْنَ وَاِنَّآ اِلٰى رَبِّنَا لَمُنْقَلِبُوْنَ

artinya: Maha suci Allah yang memudahkan ini (kendaraan) bagi kami dan tiada kami mempersekutukan bagi-Nya, dan sesungguhnya kami akan kembali kepada Tuhan kami.

Alat transportasi merupakan alat pemindahan manusia dari satu tempat ke tempat lainya. Alat transportasi sebenarnya sudah ada sejak zaman dulu. Allah SWT menciptakan berbagai macam hewan, beberapa diantaranya ialah untuk ditunggangi sebagai alat transportasi. Hal tersebut tercatat di dalam surah An-Nahl ayat 8 yang berbunyi:

وَالْخَيْلَ وَالْبِغَالَ وَالْحَمِيرَ لِتَرْكَبُوهَا وَزِينَةً ۚ وَيَخْلُقُ مَا لَا تَعْلَمُونَ ﴿٨﴾

artinya: dan (Dia telah menciptakan) kuda, bagal, dan keledai, untuk kamu tunggangi dan (menjadi) perhiasan. Allah menciptakan apa yang tidak kamu ketahui (QS. An-Nahl: 8).

Pada akhir ayat 8 pada surah An-Nahl, Allah menunjukkan bahwa adanya kemajuan dalam teknologi transportasi lainnya. Hal tersebut didukung oleh kisah Nabi Nuh As yakni Allah menciptakan kapal untuk menolong Nabi Nuh As beserta

kaumnya dari bencana banjir bandang. Hal tersebut tercatat di dalam surah yasin ayat 41-42 yang berbunyi:

وَاٰيَةٌ لَّهُمْ اَنَّا حَمَلْنَا ذُرِّيَّتَهُمْ فِى الْفُلْكِ الْمَشْحُوْنِۙ ﴿٤١﴾ وَخَلَقْنَا لَهُمْ مِّنْ مِّثْلِهٖ مَا يَرْكَبُوْنَ ﴿٤٢﴾

artinya: Dan suatu tanda (kebesaran Allah) bagi mereka adalah bahwa Kami angkut keturunan mereka dalam kapal yang penuh muatan, dan Kami ciptakan (juga) untuk mereka (angkutan lain) seperti apa yang mereka kendarai (QS. Yasin : 41-42).

Ayat 41 dan 42 pada surah Yasin menunjukkan bahwa kemajuan teknologi dalam transportasi benar adanya.

Seperti halnya kereta, pada awalnya kereta hanya dapat dijalankan menggunakan bahan bakar batubara, tetapi saat ini kereta api sudah ada yang hanya menggunakan bahan bakar listrik saja.

BAB III

METODE PENELITIAN

3.1. Jenis Penelitian

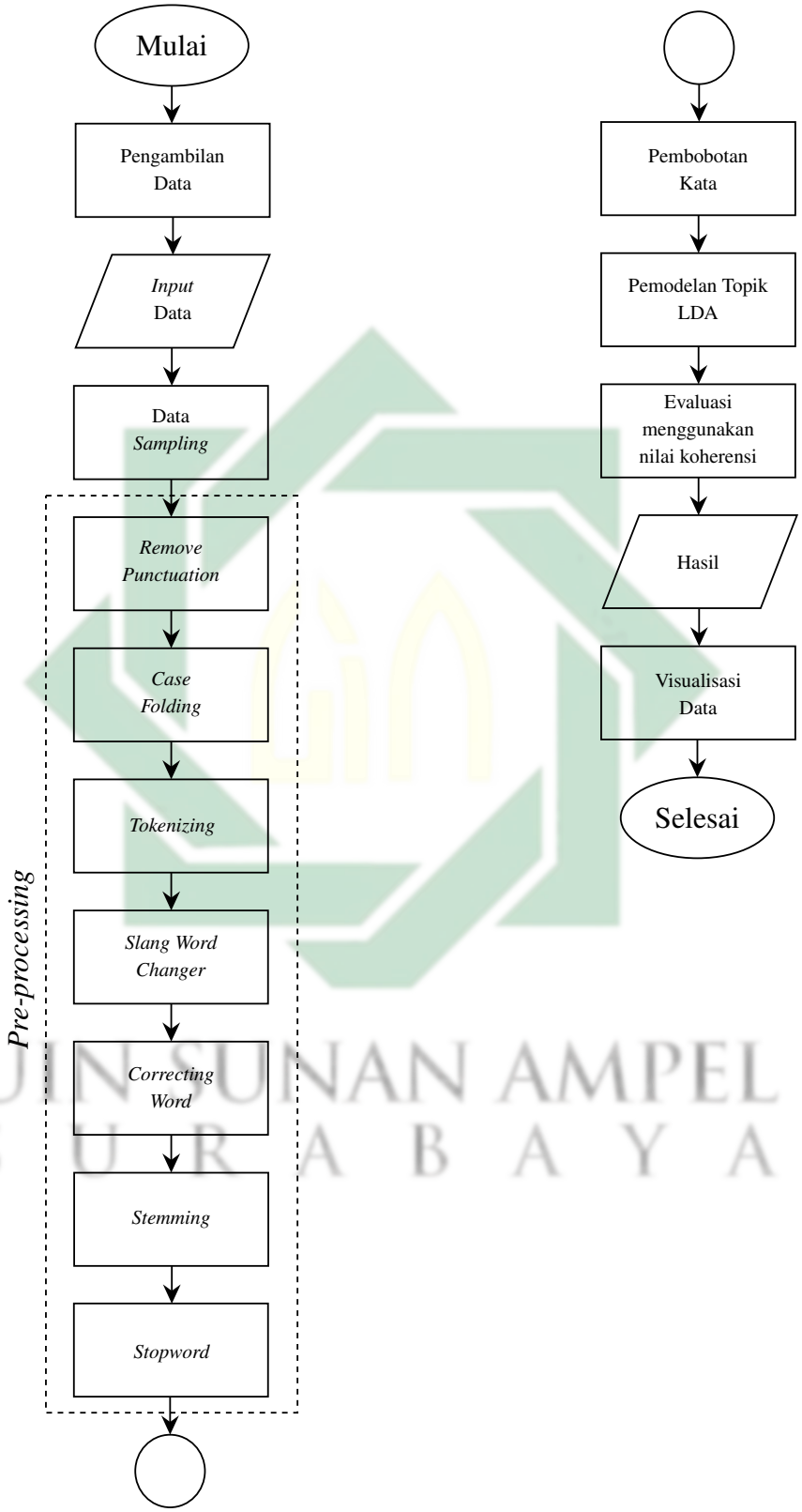
Penelitian ini adalah penelitian kuantitatif. Data yang digunakan adalah postingan pengguna twitter yang menandai akun twitter resmi dari PT Kereta Api Indonesia (Persero), lalu diolah menggunakan pendekatan matematika.

3.2. Sumber Data

Data yang digunakan dalam penelitian ini diambil menggunakan twitter API. Dataset yang digunakan ialah merupakan *tweets* para pengguna twitter dengan kata kunci pencarian ”@KAI121” yang merupakan akun twitter resmi milik PT Kereta Api Indonesia (Persero).

3.3. Tahapan Penelitian

Tahapan-tahapan dalam penyelesaian penelitian ini dijelaskan dalam diagram alir pada Gambar 3.1.



Gambar 3.1 Diagram Alir Penelitian

1. Pengambilan Data

   Pengambilan data dilakukan menggunakan Twitter API dengan menggunakan kata kunci @KAI121 dalam rentang waktu 01 Januari 2022 hingga 31 Desember 2022.

2. *Data sampling*

   Data sampling bertujuan untuk pengurangan data pada dataset aslinya tanpa mengubah bentuk struktur data. Pada penelitian metode sampling yang digunakan adalah *systematic sampling*. Data yang diambil dari dataset merupakan data dengan urutan kelipatan 100.

3. *Preprocessing*

   Pada tahap *preprocessing* dilakukan dengan beberapa langkah yakni :

   (a) *Remove Punctuation* : Pada tahapan ini, seluruh karakter yang tidak digunakan akan dihilangkan. Karakter tersebut diantaranya adalah tanda baca, nama pengguna twitter, *hashtag*, url, dan angka.

   (b) *Case Folding* : Selanjutnya, seluruh kata di dalam data akan diubah menjadi bentuk huruf kecil (*lowercase*).

   (c) *Tokenizing* : Pada tahapan ini mengubah tiap kata dalam dokumen menjadi token. Hasil dari proses ini berbentuk *array* yang berisikan token kata invidua yang terpisah.

   (d) *Slang Word Changer* : Pada tahap ini token token yang memiliki kata gaul (*slang word*) akan diubah menjadi kata baku menggunakan pedoman dataset kata gaul bahasa indonesia (Owen, 2020c). Token kata yang termasuk dalam dataset secara otomatis akan terganti dengan kata yang lebih baku.

(e) *Correcting word* : Pada tahap ini setiap huruf dalam token kata akan dicocokan dengan dataset kata dasar bahasa indonesia(Owen, 2020b). Token kata akan disamakan menggunakan token pada kata dasar melalui huruf yang sama. Token kata yang memiliki paling banyak kesamaan huruf pada suatu token kata di dalam dataset akan menjadi kata tersebut.

(f) *Stopwords* : Pada tahap ini akan menghapus kata-kata yang dianggap tidak digunakan (*stoplist*). Dalam penelitian ini terdapat 2 macam *stopwords* yang digunakan. Pertama token kata akan di sortir menggunakan dataset *stopwords* indonesia (Owen, 2020a). karena data yang digunakan berasal dari media sosial twitter, selanjutnya setiap token kata akan kembali di sortir menggunakan dataset *stopwords* twitter bahasa indonesia (Nimastiti and Julianto, 2022).

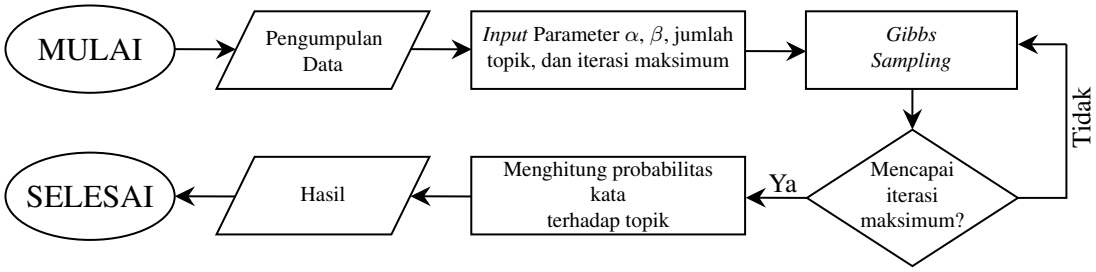
(g) *Stemming* : Pada tahap ini adalah proses menormalkan setiap token kata menjadi bentuk kata dasar.

4. Pembobotan Kata

   Pada penelitian ini pembobotan kata dilakukan dengan 2 metode yang berbeda yakni *term frequency* (TF) menggunakan persamaan 2.1 dan *term frequency inverse document frequency* (TF-IDF) menggunakan persamaan 2.2.

5. Pemodelan LDA *collapsed gibbs sampling*

   Pada tahapan ini dilakukan pemodelan topik dengan penambahan atau pengurangan jumlah topik. Proses pemodelan topik dilakukan secara berulang menggunakan *collapsed gibbs sampling* sebanyak rentang jumlah topik dan iterasi yang telah ditentukan. Alur pemodelan LDA dapat dilihat pada Gambar 3.2



Gambar 3.2 Diagram Alir LDA

Dalam menggunakan LDA, untuk mengestimasi probabilitas digunakan algoritma gibbs sampling. Langkah langkah cara kerja gibbs sampling adalah sebagai berikut

(a) Inisialisasi setiap kata yang ada pada dokumen dengan mengacak topik untuk setia kata.

(b) Menghitung probabilitas setiap kata pada topik. Dan menghitung probabilitas topik pada setiap dokumen.

(c) Menghitung probabilitas posterior menggunakan persamaan 2.11.

(d) Memilih probabilitas tertinggi dari probabilitas posterior untuk menentukan topik terbaru.

(e) Melakukan iterasi seperti langkah sebelumnya hingga mencapai iterasi maksimal

(f) Menghitung probabilitas kata terhadap topik menggunakan persamaan 2.12

(g) Menghitung probabilitas topik terhadap dokumen menggunakan persamaan 2.13

6. Evaluasi

Pada tahap ini evaluasi dilakukan dengan menggunakan nilai koherensi.

Pada penelitian ini nilai koherensi yang digunakan adalah algoritma C_v. Untuk mendapatkan nilai koherensi perlu dilakukannya 4 tahapan, diantaranya adalah segmentasi, perhitungan probabilitas, ukuran konfirmasi, dan pengumpulan nilai.

7. Visualisasi data

   Jumlah topik dengan model yang terbaik, selanjutnya tiap topik akan divisualisasikan menggunakan *word cloud*.

3.4. Uji Coba Parameter

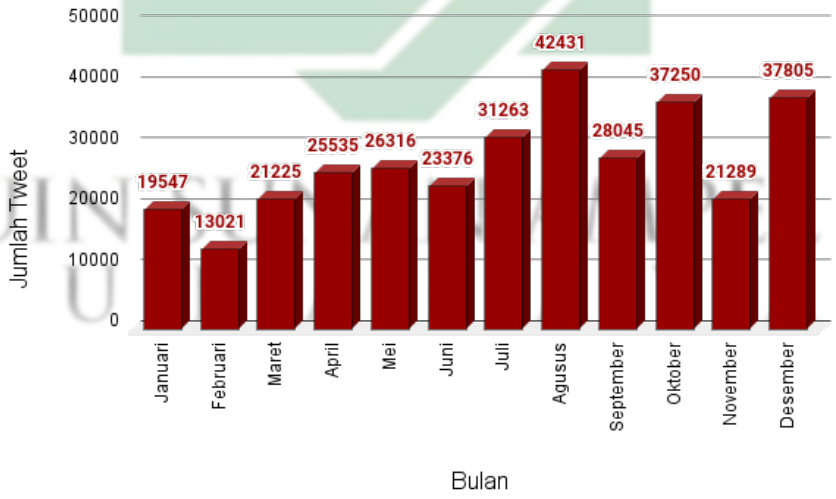
Penelitian ini melakukan beberapa uji coba parameter untuk menentukan model terbaik. Setiap uji coba dilakukan dengan menggunakan pembobotan kata yang berbeda, yakni pembobotan *term frequency* dan *term frequency inverse document frequency*. Sedangkan uji coba parameter yang dilakukan meliputi parameter dirichlet dokumen (α = 0.01, 0.05, 0.1, dan 0.5), parameter dirichlet kata (β = 0.01, 0.05, 0.1, dan 0.5), jumlah topik (K = 2, 3, 4,..., 10), dan jumlah iterasi (I = 10 dan 50).

BAB IV

HASIL DAN PEMBAHASAN

4.1. Pengambilan Data

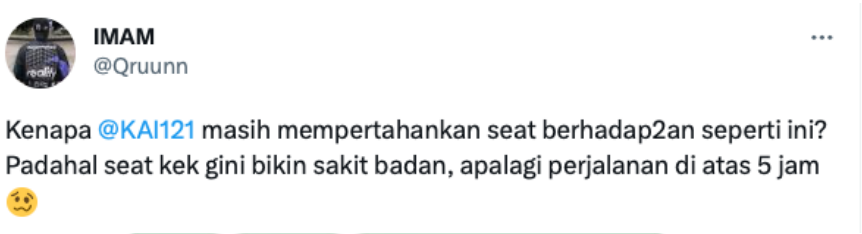
Dalam pemodelan topik opini masyarakat pengguna twitter terhadap PT Kereta Api Indonesia (Persero), data merupakan pusat objek analisis. Data yang digunakan merupakan data bertipe teks yang diambil dari tweet masyarakat pengguna twitter yang menandai akun resmi PT Kereta Api Indonesia ”@KAI121”. Pengambilan data dilakukan menggunakan TwitterAPI dengan kata kunci ”@KAI121” dari tanggal 01 Januari 2022 hingga 31 Desember 2022. Sebaran data jumlah tweet dapat dilihat pada Gambar 4.1



Gambar 4.1 Sebaran Jumlah Tweet

Hasil pengambilan data didapatkan sebanyak 327103 tweet. Dapat dilihat pada Gambar 4.1 *tweet* terbanyak pada bulan Agustus dengan jumlah *tweet*

sebanyak 42431 data. Pada bulan tersebut kata ”@kai121” menjadi viral sebabnya ada salah satu pengguna twitter yang mengungkapkan kesedihannya atas fasilitas bangku ekonomi yang kurang nyaman (Noorca, 2022). *Tweet* tersebut dapat dilihat pada Gambar 4.2

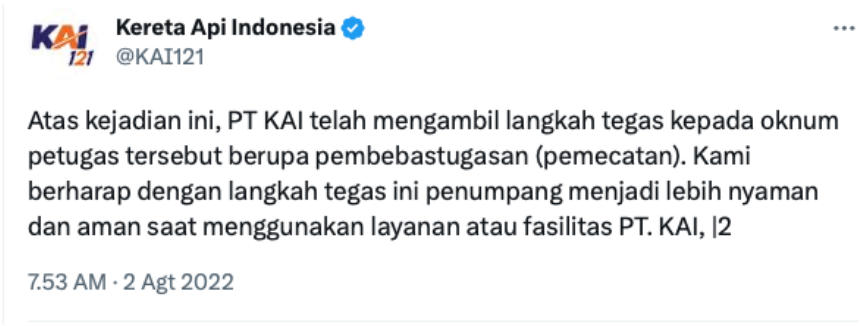


Gambar 4.2 *Tweet* Viral pada Bulan Agustus

Kejadian viral tersebut mendapatkan respon positif dari PT Kereta Api Indonesia dengan melakukan peremajaan pada setiap unit kereta ekonomi (Indonesia, 2022). Selain itu, pada bulan Agustus terdapat salah satu pengguna twitter yang melaporkan salah satu pelecehan seksual yang terjadi di salah satu stasiun yang dilakukan oleh staf pekerja PT Kereta APi Indonesia (Persero) (Fida, 2022). *Tweet* tersebut dapat dilihat pada Gambar 4.3

Gambar 4.3 *Tweet* Viral pada Bulan Agustus

Tweet tersebut menuai banyak komentar negatif mengenai PT Kereta Api Indonesia (Persero), sehingga pihak PT Kereta Api (Persero) secara langsung memberikan tindakan dengan melakukan pemecatan terhadap pegawai pelaku pelecehan seksual (Info, 2022). *Tweet* tanggapan resmi dari PT Kereta Api Indonesia (Persero) dapat dilihat pada Gambar 4.4



Gambar 4.4 *Tweet* Tanggapan PT Kereta Api Indonesia (Persero)

Dari kedua kejadian tersebut dapat diketahui bahwa twitter merupakan penghubung antara pengguna dan perusahaan PT Kereta Api Indonesia itu sendiri. Dan kedua *tweet* tersebut direspon secara positif oleh akun official dari PT Kereta Api Indonesia (Persero).

Sementara itu, pada bulan Februari paling sedikit diantara bulan yang lainnya dengan jumlah *tweet* sebanyak 13021. Rata rata banyak *tweet* opini terhadap PT Kereta Api Indonesia (Persero) setiap bulannya sekitar 27258 *tweet*. Setiap data *tweet* yang didapatkan memiliki 27 atribut. Macam macam atribut pada setiap tweet dan tipe data dapat dilihat pada Tabel 4.1.

Tabel 4.1 Atribut Data

No	Atribut	Tipe Data	No	Atribut	Tipe Data
1	Type	Object	15	Source url	Float
2	Url	Object	16	Otlink	Float
3	Date	Object	17	CountLinks	Float
4	Content	Object	18	Media	Float
5	Rendered Content	Object	19	RetweetedTweet	Object
6	Id	Int	20	QuotedTweet	Object
7	User	Object	21	InReplyToTweetid	Float
8	ReplyCount	Int	22	InReplyToUser	Int
9	RetweetCount	Int	23	MentionedUsers	Object
10	LikeCount	Int	24	Coordinates	Float
11	QuoteCount	Int	25	Place	Float
12	Converstationid	Int	26	Hastaghs	Object
13	Lang	Object	27	Cashtags	Object
14	Source	Object			

Contoh tweet hasil pengambilan data pada twitterAPI dengan kata pencarian ”@KAI121” dapat dilihat pada Tabel 4.2. Hasil yang ditampilkan hanya bagian yang penting, yaitu pada atribut *id* dan *content*.

Tabel 4.2 Sampel Data Tweet Opini Terhadap PT Kereta Api Indonesia (Persero)

id	Content
1572389022760310000	@KAI121 @bart_the_13th knp klo naik kereta ke arah cicalengka selalu diem lama di stasiun gedebage?
1572389022760310000	Gila sih kereta lokal cicalengka-bandung telatnya ga kira-kira, mana udah malem pula 🤡🤡 @KAI121

4.2. *Data Sampling*

Pada penelitian ini, dilakukan data sampling karena banyaknya data yang diperoleh dapat menghambat cara kerja LDA. Teknik data sampling yang digunakan dalam penelitian ini adalah *systematic sampling* dengan nilai subjek pertama adalah 100. Lalu subjek berikutnya yakni berkelipatan 100.

Dari 327104 tweet yang di peroleh, setelah dilakukannya proses *systematic sampling* menyisakan sejumlah 3271 tweet opini terhadap PT Kereta Api Indonesia (Persero).

4.3. *Preprocessing*

Tahap preprocessing data merupakan tahap yang krusial. Karena pada tahap ini menyiapkan data untuk menormalkan dataset agar mudah menentukan analisis aspek LDA. Pada penelitian ini, *preprocessing* data yang telah dilakukan adalah *remove punctuation*, *case folding*, *tokenizing*, *slang word changer*, *correcting typo*, *stemming*, *stopword*.

4.3.1. *Remove Punctuation*

Pada tahap ini menghapus seluruh karakter yang tidak digunakan dalam penelitian seperti angka, tanda baca, *hashtag*, *username*, dan *emoticon* yang ada di dalam data. Hasil pada tahap ini dapat dilihat pada Tabel 4.3.

Tabel 4.3 Proses *Remove Punctuation*

Sebelum *Remove Punctuation*	Sesudah *Remove Punctuation*
@KAI121 @bart_the_13th knp klo naik kereta ke arah cicalengka selalu diem lama di stasiun gedebage?	knp klo naik kereta ke arah cicalengka selalu diem lama di stasiun gedebage
Gila sih kereta lokal cicalengka-bandung telatnya ga kira kira, mana udah malem pula @KAI121	gila sih kereta lokal cicalengka bandung telatnya ga kira kira mana udah malem pula

Seperti pada Tabel 4.3,kata yang berwarna merah merupakan *noise* pada data tersebut sehingga perlu dihapus agar tidak mengganggu jalannya proses analisis dalam LDA.

4.3.2. *Case Folding*

Case folding merupakan tahap merubah seluruh kosakata di dalam setiap data menjadi huruf kecil (*lower case*). Contoh dari tahap ini adalah Jakarta berubah menjadi jakarta, ATM berubah menjadi atm, dan lain sebagainya. Tahap *case folding* dapat dilihat pada Tabel 4.4.

Tabel 4.4 Proses *Case Folding*

Sebelum *Case Folding*	Sesudah *Case Folding*
knp klo naik kereta ke arah cicalengka selalu diem lama di stasiun gedebage	knp klo naik kereta ke arah cicalengka selalu diem lama di stasiun gedebage
gila sih kereta lokal cicalengka bandung telatnya ga kira kira mana udah malem pula	gila sih kereta lokal cicalengka bandung telatnya ga kira kira mana udah malem pula

4.3.3. *Tokenizing*

Berikutnya, pada tahap *tokenizing* mengubah data dimana data tersebut berbentuk kalimat menjadi potongan individu kata. Proses ini dapat mempermudah pada tahap berikutnya. Tahap *tokenizing* dapat dilihat pada Tabel 4.5

Tabel 4.5 Proses *Tokenizing*

Sebelum *Tokenizing*	Sesudah *Tokenizing*
knp klo naik kereta ke arah cicalengka selalu diem lama di stasiun gedebage	['knp', 'klo', 'naik', 'kereta', 'ke', 'arah', 'cicalengka', 'selalu', 'diem', 'lama', 'di', 'stasiun', 'gedebage']
gila sih kereta lokal cicalengka bandung telatnya ga kira kira mana udah malem pula	['gila', 'sih', 'kereta', 'lokal', 'cicalengka', 'bandung', 'telatnya', 'ga', 'kira', 'kira', 'mana', 'udah', 'malem', 'pula']

Dapat dilihat pada Tabel 4.5 data yang masih berupa kalimat utuh menjadi satu *list* kata kata. *List* tersebutlah yang akan menjadi input pada tahap berikutnya.

4.3.4. *Slang Word Changer*

Data yang sudah berupa beberapa *list* token kata, selanjutnya diproses ke dalam tahap *slang word changer*. Pada tahap ini, setiap token kata pada data yang mengandung kata gaul *slang word* akan diubah menjadi kata yang baku menurut KBBI. Proses *slang word changer* dapat dilihat pada Tabel 4.6

Tabel 4.6 Proses *Slang Word Changer*

Sebelum *Slang Word Changer*	Sesudah *Slang Word Changer* `
['knp', 'klo', 'naik', 'kereta', 'ke', 'arah', 'cicalengka', 'selalu', 'diem', 'lama', 'di', 'stasiun', 'gedebage']	['kenapa', 'kalau', 'naik', 'kereta', 'ke', 'arah', 'cicalengka', 'selalu', 'diem', 'lama', 'di', 'stasiun', 'gedebage']
['gila', 'sih', 'kereta', 'lokal', 'cicalengka', 'bandung', 'telatnya', 'ga', 'kira', 'kira', 'mana', 'udah', 'malem', 'pula']	['gila', 'sih', 'kereta', 'lokal', 'cicalengka', 'bandung', 'telatnya', 'tidak', 'kira', 'kira', 'mana', 'sudah', 'malem', 'pula']

Dapat dilihat pada Tabel 4.6 pada kata ”knp” berubah menjadi kata ”kenapa”, kata ”klo” berubah menjadi kata ”kalau”, dan kata ”ga” berubah menjadi bentuk baku kata ”tidak”.

4.3.5. *Correcting Typo*

Pada tahap ini, token token kata yang telah diubah ke dalam bahasa baku kemudian dilakukan pengecekan kesalahan kata. Pengecekan kesalahan kata (*correcting typo*) dilakukan menggunakan dataset kata dasar dengan memperhatikan jumlah huruf yang paling banyak benarnya. Hasil pada tahap *correcting typo* dapat dilihat pada Tabel 4.7.

Tabel 4.7 Proses *Correcting Typo*

Sebelum *Correcting Typo*	Sesudah *Correcting Typo*
['kenapa', 'kalau', 'naik', 'kereta', 'ke', 'arah', 'cicalengka', 'selalu', 'diem', 'lama', 'di', 'stasiun', 'gedebage']	['kenapa', 'kalau', 'naik', 'kereta', 'ke', 'arah', 'cicalengka', 'selalu', 'diam', 'lama', 'di', 'stasiun', 'gedebage']
['gila', 'sih', 'kereta', 'lokal', 'cicalengka', 'bandung', 'telatnya', 'tidak', 'kira', 'kira', 'mana', 'sudah', 'malem', 'pula']	['gila', 'sih', 'kereta', 'lokal', 'cicalengka', 'bandung', 'telatnya', 'tidak', 'kira', 'kira', 'mana', 'sudah', 'malam', 'pula']

Pada data sampel yang telah dilampirkan terdapat beberapa penulisan kata yang salah seperti pada Tabel 4.7. Token kata ”malem” berubah menjadi ”malam”.

4.3.6. *Stemming*

Tahap berikutnya pada proses *preprocessing* adalah tahap *stemming*. Tahapan ini dapat merubah token kata menjadi kata dasar. Salah satu bentuk merubah token kata menjadi kata dasar ialah dengan menghilangkan kata yang memiliki imbuhan seperti, ”permohonan” berubah menjadi ”mohon”, ”petugas” berubah menjadi ”tugas”, dan lain sebagainya. Hasil proses *stemming* dapat dilihat pada Tabel 4.8

Tabel 4.8 Proses *Stemming*

Sebelum *Stemming*	Seusdah *Stemming*
['kenapa', 'kalau', 'naik', 'kereta', 'ke', 'arah', 'cicalengka', 'selalu', 'diam', 'lama', 'di', 'stasiun', 'gedebage']	['kenapa', 'kalau', 'naik', 'kereta', 'ke', 'arah', 'cicalengka', 'selalu', 'diam', 'lama', 'di', 'stasiun', 'gedebage']
['gila', 'sih', 'kereta', 'lokal', 'cicalengka', 'bandung', 'telatnya', 'tidak', 'kira', 'kira', 'mana', 'sudah', 'malam', 'pula']	['gila', 'sih', 'kereta', 'lokal', 'cicalengka', 'bandung', 'telat', 'tidak', 'kira', 'kira', 'mana', 'sudah', 'malam', 'pula']

Dapat dilihat pada Tabel 4.8, token kata ”telatnya” berubah menjadi kata dasar ”telat” dan menghilangkan imbuhan kata ”nya” pada data.

4.3.7. Stopword

Tahap terakhir pada *preprocessing* adalah tahap *stopwords*. Pada tahap ini token kata pada setiap data yang tidak memiliki arti akan dihilangkan. Kata kata yang tidak memiliki arti tersebut merupakan kata depan, kata sapaan, kata pertanyaan dan lain sebagainya. Hasil tahap *stopwords* dapat dilihat pada Tabel 4.9.

Tabel 4.9 Proses *Stopword*

Sebelum *Stopword*	Sesudah *Stopword*
['kenapa', 'kalau', 'naik', 'kereta', 'ke', 'arah', 'cicalengka', 'selalu', 'diam', 'lama', 'di', 'stasiun', 'gedebage']	['kereta', 'arah', 'cicalengka', 'diam', 'lama', 'stasiun', 'gedebage']
['gila', 'sih', 'kereta', 'lokal', 'cicalengka', 'bandung', 'tepat', 'tidak', 'kira', 'kira', 'mana', 'sudah', 'malam', 'pula']	['gila', 'kereta', 'lokal', 'cicalengka', 'bandung']

Pada Tabel 4.9 token kata yang berwarna merah merupakan contoh token kata yang termasuk dalam *stoplist* di dalam *stopword*. Kata tersebut harus dihilangkan karena tidak digunakan dalam penelitian ini.

4.4. Pembobotan Kata

Setelah tahap *preprocessing*, jumlah data yang pada awalnya berjumlah 3271 tweet setelah melewati tahap *preprocessing* menghasilkan 2915 data dengan token kata sebanyak 4014 token. Langkah berikutnya pada penelitian ini adalah pembobotan kata.

Pembobotan kata bertujuan untuk mengubah data yang berupa teks menjadi angka. Langkah awal dalam pembobotan kata yakni menginisialisasikan kata ke dalam angka. Langkah tersebut dapat dilihat pada tabel 4.10.

Tabel 4.10 Inisialisasi Token Kata

Kata	Inisialisasi
coba	1
tim	2
ayonaik	3
. . .	. . .
kendor	4012
penumpang	4013
pembatalan	4014

Dapat dilihat pada Tabel 4.10 penginisialisasi kata dimulai pada nilai 1 sampai 4014, penginisialisasi kata dimulai dari kata pertama pada dokumen pertama hingga token kata pada korpus habis. Setelah penginisialisasi dilakukan selanjutnya menghitung bobot kata *term frequency*.

4.4.1. *Term Frequency*

(TF) Pembobotan *term frequency* (TF) merupakan pembobotan dengan menghitung jumlah kemunculan token kata dalam suatu dokumen. Hasil

pembobotan kata *term frequency* dapat dilihat pada Tabel 4.11.

Tabel 4.11 Hasil Pembobotan Kata *Term Frequency*

	1	2	3	4	5	6	⋯	4013	4014
dok 1	1	1	0	0	0	0	⋯	0	0
dok 2	1	1	1	1	1	1	⋯	0	0
dok 3	0	0	0	0	0	0	⋯	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮
dok 2913	0	0	0	0	0	0	⋯	0	0
dok 2914	0	0	0	0	0	0	⋯	0	0
dok 2915	0	0	0	0	0	0	⋯	1	1

Dapat dilihat pada Tabel 4.11 diatas merupakan hasil dari pembobotan *term frequency* yang merupakan jumlah frekuensi kemunculan kata pada dokumen. Jika dilihat pada Tabel 4.11 kata pertama pada dokumen 1 yakni ”coba” bernilai 1 berarti kata tersebut muncul sebanyak satu kali pada dokumen 1. Begitu pula jika suatu kata tidak muncul pada dokumen maka nilai bobotnya 0. Tabel 4.11 jika berbentuk matriks akan sebagai berikut :

$$W = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 1 \end{bmatrix}_{2915 \times 4014}$$

4.4.2. *Term Frequency Inverse Document Frequency*

(TF-IDF) Setelah mendapatkan nilai bobot *term frequency*, langkah selanjutnya adalah menghitung bobot *term frequency inverse document frequency* TF-IDF. Langkah ini, hasil dari perkalian TF dengan IDF.

Sebagai contoh, diambil pada dokumen pertama kata pertama yakni ”coba”. Pada Tabel 4.11 nilai TF kata ”coba” pada dokumen 1 sebesar 1, kemudian untuk menghitung nilai IDF diperlukan *document frequency* yang merupakan jumlah frekuensi dokumen yang terdapat kata ”coba” yakni sebanyak 41. Lalu dalam menghitung nilai IDF menggunakan persamaan 2.2 dengan jumlah dokumen sebanyak 2915 :

$$
\begin{aligned}
W_{w,v} &= tf_{w,v} \times \log\left(\frac{N}{df_0}\right) \\
W_{1,1} &= 1 \times \log\left(\frac{2915}{41}\right) \\
&= 1 \times \log(62.0212) \\
&= 1.852
\end{aligned}
$$

Didapatkan bobot dari kata ”coba” pada dokumen 1 sebesar 1.8518. Hasil perhitungan dari TF-IDF dapat dilihat pada Tabel 4.12.

Tabel 4.12 Hasil Pembobotan TF-IDF

	1	2	3	4	5	6	⋯	4013	4014
dok 1	1.852	2.686	o	0	0	0	⋯	0	0
dok 2	0	0	2.465	1.908	3.465	1.094	⋯	0	0
dok 3	0	0	0	0	0	0	⋯	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮
dok 2913	0	0	0	0	0	0	⋯	0	0
dok 2914	0	0	0	0	0	0	⋯	0	0
dok 2915	0	0	0	0	0	0	⋯	3.465	3.465

Dapat dilihat pada Tabel 4.12 setiap token cenderung memiliki bobot yang berbeda. Hal tersebut terjadi karena token token pada korpus memiliki frekuensi terhadap dokumen yang berbeda - beda juga. Tabel 4.12 jika berbentuk matriks akan sebagai berikut :

$$W = \begin{bmatrix} 1.852 & 2.687 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 2.465 & 1.908 & 3.465 & 1.094 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 3.465 & 3.465 \end{bmatrix}_{2915 \times 4014}$$

4.5. *Latent Dirichlet Allocation*

Dalam memodelkan topik, *latent dirichlet allocation* membutuhkan beberapa parameter. Parameter tersebut digunakan berdasarkan penelitian sebelumnya, nilai yang pas untuk parameter α adalah di rentang 0.1 hingga 1. Sedangkan parameter β adalah di rentang 0.01 hingga 1 (Darussalam and Arief, 2017) (Nugroho et al., 2021). Dari beberapa penelitian yang pernah dilakukan,

jumlah iterasi 10 merupakan jumlah yang paling optimal untuk pemodelan topik LDA (Nugroho et al., 2021) (Darussalam and Arief, 2017).Oleh karena itu, parameter yang digunakan pada penelitian ini adalah sebagai berikut :

1. α = 0.01, 0.05, 0.1, dan 0.5
2. β = 0.01, 0.05, 0.1, dan 0.5
3. K = 2, 3, 4,..., 10
4. Jumlah iterasi = 10 dan 50

Setelah menentukan parameter langkah berikutnya adalah menetapkan topik untuk setiap kata dalam dokumen secara acak. penentuan topik tersebut dimasukkan ke dalam matriks z dimana jumlah matriks z sebanyak jumlah topik yang dilakukan. Matriks ini berisikan nilai biner yang artinya nilai 0 untuk kata yang tidak termasuk dalam topik tersebut, sedangkan nilai 1 untuk kata yang termasuk dalam topik tersebut. Sebagai contoh digunakan jumlah topik sebanyak 2 sehingga matriks z dapat dilihat sebagai berikut

$$Z_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}_{2915 \times 4014} \tag{4.1}$$

$$Z_2 = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{2915 \times 4014} \tag{4.2}$$

Kemudian untuk token kata unik dalam dokumen (W) didapatkan dari hasil pembobotan kata. Karena penelitian ini menggunakan jenis pembobotan yang berbeda, maka matriks W terdapat 2 macam yang akan dihitung secara berbeda lalu hasilnya akan dibandingkan.

4.5.1. *Collapsed Gibbs Sampling*

Setelah melakukan pembobotan kata, untuk mengestimasi tiap topik pada token digunakan algoritma *gibbs sampling*. Di dalam gibbs sampling diperlukan dua matriks yang digunakan diantaranya adalah matriks representasi dari jumlah kata pada topik di seluruh korpus (TK) dan matriks representasi dari jumlah kata pada dokumen yang ditetapkan pada topik (KD).

Matriks TK merupakan hasil penjumlahan tiap kolom dari perkalian per elemen dari matriks Z dengan matriks W. Sebagai contoh dengan jumlah topik 2

maka matriks TK berukuran 2×4014. Baris pertama merupakan representasi dari jumlah kata dalam korpus pada topik 1, atau jumlah tiap kolom pada persamaan 4.5.1.. Sementara baris kedua merupakan representasi dari jumlah kata dalam korpus pada topik 2, atau jumlah tiap kolom pada persamaan 4.2. Matriks TK dapat dilihat sebagai berikut :

$$TK = \begin{bmatrix} 2 & 15 & \cdots & 1 & 0 \\ 4 & 25 & \cdots & 0 & 1 \end{bmatrix}_{2\times 4014}$$

Selanjutnya pembuatan matriks n, ialah dengan menjumlah frekuensi setiap topik pada tiap dokumen di dalam korpus atau dengan kata lain matriks KD merupakan jumlah tiap baris pada matriks Z. Sebagai contoh yang telah dilakukan dengan jumlah topik 2, matriks KD berukuran 2195×2. kolom pertama merupakan representasi dari jumlah topik 1 pada tiap dokumen, atau jumlah tiap baris pada persamaan 4.5.1.. Sementara kolom kedua merupakan representasi dari jumlah topik 2 pada tiap dokumen, atau jumlah tiap baris pada persamaan 4.2. Matriks KD dapat dilihat sebagai berikut :

$$KD = \begin{bmatrix} 1 & 1 \\ 4 & 5 \\ \vdots & \vdots \\ 9 & 4 \\ 5 & 6 \end{bmatrix}_{2195\times 2}$$

Karena topik awal ditentukan secara acak, setelah kedua matriks tersebut di bentuk langkah berikutnya yakni mencari topik baru untuk setiap token. Perhitungan topik baru tersebut dengan menggunakan persamaan 2.11. Sebagai

contoh akan mencari topik baru pada dokumen 1 kata pertama yakni ”coba” dengan menggunakan bobot kata TF. Pada penentuan acak token kata pada tersebut berada pada topik 1. Penentuan topik baru menggunakan gibbs sampling sebagai berikut, diketahui :

$$\alpha = 0.1 \qquad \beta = 0.01$$

$$TK_{1,1} = 2 - 1 = 1 \qquad TK_{2,1} = 4$$

$$\sum_{v=1}^{V} TK_{1,v} = 11188 - 1 \qquad \sum_{v=1}^{V} TK_{2,v} = 11213$$

$$\sum_{k=1}^{K} KD_{1,k} = 2 - 1 = 1 \qquad KD_{1,1} = 1 - 1 = 0$$

$$KD_{1,2} = 1 \qquad N = 4014$$

$$K = 2$$

Sehingga nilai $p(z_1)$:

$$p(z_i = k|w_i = v, d, \alpha, \beta) = \frac{TK_{1,1} + \beta}{\sum_{v=1}^{V} TK_{1_v} + \beta V} \frac{KD_{1,1} + \alpha}{\sum_{k=1}^{K} KD_{d_1} + \alpha K}$$

$$p(z_i = 1|w_i = 1, 1, \alpha, \beta) = \frac{1 + 0.01}{11187 + 0.01(4014)} \frac{0 + 0.1}{1 + 0.1(2)}$$

$$= \frac{1.01}{11227,14} \frac{0.1}{1.2}$$

$$= 10^{-9} \times 7496$$

Kemudian menghitung nilai $p(z_2)$ sebagai berikut:

$$p(z_i = k|w_i = v, d, \alpha, \beta) = \frac{TK_{2,1} + \beta}{\sum_{v=1}^{V} TK_{2_v} + \beta V} \frac{KD_{1,2} + \alpha}{\sum_{k=1}^{K} KD_{d_2} + \alpha K}$$

$$p(z_i = 2|w_i = 1, 1, \alpha, \beta) = \frac{4 + 0.01}{11213 + 0.01(4014)} \frac{1 + 0.1}{1 + 0.1(2)}$$

$$= \frac{4.01}{11253,14} \frac{1.1}{1.2}$$

$$= 10^{-7} \times 4703$$

Berdasarkan perhitungan nilai $p(z)$ diatas, nilai $p(z_2)$ lebih besar daripada nilai $p(z_1)$. Oleh karena itu, topik baru untuk kata pertama dalam dokumen 1 yakni berada pada topik 2. Perhitungan topik baru dilakukan secara berulang hingga seluruh token kata pada setiap dokumen dan berulang lagi sebanyak jumlah iterasi yang diinginkan. Sehingga matriks Z_1 dan Z_2 menjadi sebagai berikut:

$$Z_1 = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{2915 \times 4014}$$

$$Z_2 = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}_{2915 \times 4014}$$

Karena elemen pada matrik Z_K nya berubah, maka matriks TK dan matriks

KD pun berbeda. Sehingga kedua matriks tersebut menjadi sebagai berikut :

$$TK = \begin{bmatrix} 1 & 22 & \cdots & 0 & 0 \\ 5 & 28 & \cdots & 1 & 1 \end{bmatrix}_{2\times 4014}$$

$$KD = \begin{bmatrix} 0 & 2 \\ 7 & 2 \\ \vdots & \vdots \\ 2 & 2 \\ 8 & 3 \end{bmatrix}_{2915\times 2}$$

Setelah ditemukan topik baru pada seluruh token di dalam korpus, selanjutnya adalah menghitung probabilitas setiap kata pada topik. Perhitungan probabilitas kata pada topik menggunakan persamaan 2.12. Sebagai contoh akan menghitung probabilitas kata pertama pada dokumen 1 yakni ”coba” pada topik 2, diketahui :

$$\beta = 0.01 \qquad TK_{2,1} = 5$$

$$N = 4014 \qquad \sum_{v=1}^{V} TK_{1,v} = 11188$$

Sehingga didapatkan nilai probabilitas kata terhadap topik sebagai berikut:

$$\begin{aligned} \Phi^{11} &= \frac{W_{k_1 v_i} + \beta}{\sum_{v=1}^{V} W_{k_1 v_i} + \beta V} \\ &= \frac{5 + 0.01}{11088 + 0,01(4014)} \\ &= \frac{5.01}{11128.14} \\ &= 10^{-8} \times 4502 \end{aligned}$$

Sehingga didapatkan nilai probabilitas kata ”coba” pada dokumen 1 terhadap topik 2 sebesar $10^{-8} \times 4502$. Perhitungan ini diulang hingga token kata pada tiap topik habis.

Selanjutnya untuk mengetahui dokumen mana saja yang mengandung suatu topik tertentu perlu menghitung probabilitas topik di dalam suatu dokumen menggunakan persamaan 2.13. Sebagai contoh menghitung probabilitas topik 1 pada dokumen 2, diketahui:

$$\alpha = 0.1 \qquad KD_{2,1} = 7$$

$$K = 2 \qquad \sum_{k=1}^{K} KD_{2,k} = 9$$

Sehingga didapatkan nilai probabilitas kata terhadap topik sebagai berikut:

$$\begin{aligned} \Theta^{2,1} &= \frac{KD_{2,1} + \alpha}{\sum_{k=1}^{K} DK_{2,k} + \alpha K} \\ &= \frac{7 + 0.1}{9 + 0.1(2)} \\ &= \frac{7.1}{11.1} \\ &= 0.78 \end{aligned}$$

Sehingga didapatkan nilai probabilitas topik 1 pada dokumen ke-2 yakni sebesar 0.78. Nilai tersebut akan menjadi distribusi topik pada dokumen. Perhitungan tersebut dilakukan secara berulang sejumlah seluruh topik pada setiap dokumen.

4.6. Evaluasi Model

Pemodelan topik yang optimal didapatkan dengan melakukan beberapa kali uji coba dengan beberapa parameter yang telah ditentukan dan jumlah topik yang berbeda, agar tepat dalam melakukan analisis. Percobaan tersebut masing masing diuji menggunakan pembobotan kata TF dan TF-IDF. Hasil dalam percobaan menggunakan pembobotan TF dan TF-IDF adalah sebagai berikut:

4.6.1. LDA Menggunakan *Term Frequency*

Pemodelan topik LDA menggunakan pembobotan kata TF dengan menggunakan uji coba parameter yang telah ditentukan, dengan total 32 uji coba yang ditetapkan pada tiap topik. Hasil nilai koherensi seluruh uji coba pemodelan topik LDA dengan pembobotan kata TF dapat dilihat pada Tabel 4.13.

Dapat dilihat pada Tabel 4.13 hasil nilai koherensi dari parameter yang telah ditentukan memiliki nilai yang berfluktasi. Artinya adalah nilai tersebut tidak stabil dan tidak dipengaruhi oleh parameter yang telah ditentukan. Pada analisis lebih lanjut perlu dilihat dari rata rata nilai koherensi tiap parameter pada grafik yang dapat dilihat pada Gambar 4.5

(a) Parameter α **(b) Parameter β**

Gambar 4.5 Rata-rata Nilai Koherensi Tiap Parameter Menggunakan *Term Frequency*

Dari Gambar 4.5a diketahui bahwa rata-rata nilai koherensi α terbaik pada

jumlah iterasi 10 dan 50 berbeda. Pada jumlah iterasi 10 didapatkan nilai koherensi terbaik pada parameter α sebesar 0.01, sedangkan pada jumlah iterasi 50 nilai koherensi terbaik pada parameter α sebesar 0.05 dan 0.1. Rata-rata nilai koherensi terkecil pada jumlah iterasi 10 dan 50 keduanya sama yakni pada parameter α sebesar 0.5.

Sementara pada Gambar 4.5b diketahui bahwa rata-rata nilai koherensi β terbaik pada jumlah iterasi 10 dan 50 berbeda. Pada jumlah iterasi 10 didapatkan nilai koherensi terbaik pada parameter β sebesar 0.5, sedangkan pada jumlah iterasi 50 nilai koherensi terbaik pada parameter β sebesar 0.05. Rata-rata nilai koherensi terkecil pada jumlah iterasi 10 pada parameter β sebesar 0.05. Dan rata-rata nilai koherensi terkecil pada jumlah iterasi 50 pada parameter β sebesar 0.01

Model terbaik pada seluruh uji coba menggunakan pembobotan kata TF pada Tabel 4.13 didapatkan pada jumlah topik 10 dengan nilai parameter α sebesar 0.05, β sebesar 0.01 dengan jumlah iterasi sebanyak 50 kali mendapatkan nilai koherensi sebesar 0.685. Sedangkan model dengan nilai koherensi terendah yakni pada jumlah topik sebanyak 2 dengan parameter α sebesar 0.5, parameter β sebesar 0.01, dan jumlah iterasi sebanyak 50 kali mendapatkan nilai koherensi sebesar 0.301.

Tabel 4.13 Nilai Koherensi Pemodelan Topik LDA Menggunakan Pembobotan *Term Frequency*

α	β	Iterasi	Nilai Koherensi pada Jumlah Topik									Rata-rata
			2	3	4	5	6	7	8	9	10	
0.01	0.01	10	0.528	0.483	0.479	0.468	0.508	0.484	0.505	0.483	0.482	0.491
		50	0.445	0.478	0.391	0.466	0.415	0.453	0.444	0.421	0.462	0.442
	0.05	10	0.430	0.469	0.527	0.507	0.436	0.492	0.432	0.509	0.488	0.477
		50	0.450	0.554	0.562	0.529	0.512	0.513	0.527	0.502	0.509	0.518
	0.1	10	0.505	0.443	0.516	0.492	0.585	0.448	0.498	0.511	0.515	0.501
		50	0.489	0.522	0.451	0.534	0.484	0.479	0.508	0.465	0.503	0.493
	0.5	10	0.377	0.523	0.579	0.570	0.552	0.506	0.463	0.456	0.487	0.501
		50	0.404	0.517	0.514	0.327	0.357	0.378	0.370	0.390	0.390	0.405
0.05	0.01	10	0.479	0.531	0.529	0.509	0.520	0.462	0.485	0.456	0.501	0.497
		50	**0.401**	**0.474**	**0.538**	**0.530**	**0.489**	**0.476**	**0.485**	**0.480**	**0.685**	**0.506**
	0.05	10	0.420	0.458	0.543	0.475	0.527	0.502	0.496	0.498	0.426	0.483
		50	0.462	0.485	0.456	0.501	0.516	0.492	0.516	0.492	0.549	0.497
	0.1	10	0.435	0.538	0.464	0.423	0.508	0.468	0.486	0.488	0.488	0.478
		50	0.482	0.488	0.517	0.470	0.499	0.453	0.512	0.435	0.529	0.487
	0.5	10	0.510	0.511	0.491	0.483	0.476	0.487	0.476	0.493	0.503	0.492
		50	0.478	0.470	0.476	0.472	0.484	0.465	0.470	0.430	0.463	0.467
0.1	0.01	10	0.496	0.445	0.437	0.460	0.461	0.492	0.521	0.479	0.429	0.469
		50	0.516	0.562	0.441	0.486	0.422	0.492	0.521	0.479	0.429	0.483
	0.05	10	0.466	0.514	0.526	0.529	0.426	0.518	0.471	0.478	0.519	0.494
		50	0.499	0.453	0.512	0.435	0.529	0.530	0.514	0.502	0.480	0.495
	0.1	10	0.520	0.473	0.459	0.484	0.528	0.482	0.488	0.517	0.470	0.491
		50	0.470	0.394	0.582	0.353	0.550	0.410	0.590	0.480	0.470	0.478
	0.5	10	0.563	0.507	0.471	0.474	0.536	0.583	0.472	0.469	0.495	0.508
		50	0.476	0.513	0.469	0.551	0.471	0.531	0.537	0.476	0.476	0.500
0.5	0.01	10	0.467	0.528	0.513	0.526	0.440	0.482	0.480	0.499	0.429	0.485
		50	0.301	0.376	0.496	0.410	0.480	0.455	0.491	0.431	0.378	0.424
	0.05	10	0.455	0.430	0.525	0.496	0.489	0.538	0.414	0.480	0.523	0.483
		50	0.399	0.442	0.530	0.479	0.550	0.410	0.590	0.376	0.490	0.474
	0.1	10	0.425	0.482	0.474	0.462	0.467	0.481	0.503	0.455	0.480	0.470
		50	0.449	0.502	0.485	0.524	0.452	0.467	0.521	0.515	0.496	0.490
	0.5	10	0.599	0.514	0.457	0.493	0.494	0.528	0.491	0.492	0.492	0.507
		50	0.457	0.540	0.518	0.510	0.570	0.442	0.468	0.494	0.475	0.497

Tabel 4.14 Nilai Koherensi Tiap Topik Pembobotan *Term Frequency*

Id Topik	Nilai Koherensi
1	0.743
2	0.598
3	0.634
4	0.733
5	0.672
6	0.642
7	0.719
8	0.731
9	0.662
10	0.718

Dapat diketahui pada tabel 4.14 nilai topik pada model terbaik LDA pada pembobotan Kata TF memiliki nilai yang seragam, Hal tersebut karena pembobotan TF cenderung memiliki nilai bobot yang sama sehingga probabilitas kata yang didapatkan pun juga cenderung sama.

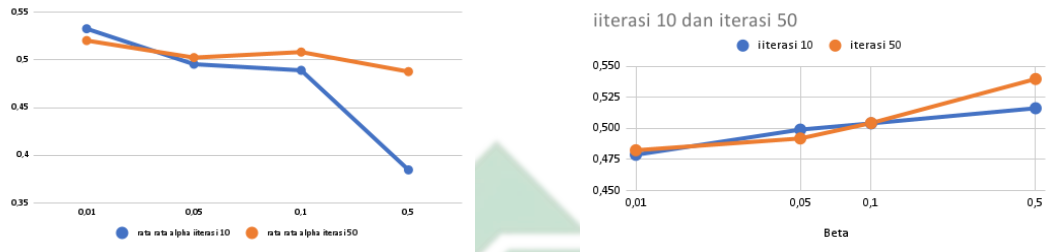
4.6.2. LDA Menggunakan *Term Frequency Inverese Document Frequency*

Pemodelan topik LDA menggunakan pembobotan kata TF-IDF dengan menggunakan uji coba parameter yang telah ditentukan, dengan total 32 uji coba yang ditetapkan pada tiap topik. Hasil nilai koherensi seluruh uji coba pemodelan topik LDA dengan pembobotan kata TF-IDF dapat dilihat pada Tabel 4.15.

Tabel 4.15 Nilai Koherensi Pemodelan Topik LDA Menggunakan Pembobotan *Term Frequency Inverse Document Frequency*

α	β	Iterasi	Nilai Koherensi pada Jumlah Topik									Rata rata
			2	3	4	5	6	7	8	9	10	
0.01	0.01	10	0.466	0.551	0.446	0.608	0.332	0.464	0.496	0.470	0.470	0.478
		50	0.506	0.429	0.474	0.510	0.590	0.604	0.336	0.447	0.483	0.487
	0.05	10	0.547	0.586	0.591	0.384	0.410	0.456	0.536	0.724	0.532	0.530
		50	0.453	0.521	0.527	0.498	0.477	0.526	0.495	0.527	0.508	0.504
	0.1	10	0.485	0.498	0.867	0.521	0.508	0.493	0.423	0.627	0.568	0.554
		50	0.588	0.532	0.469	0.475	0.504	0.546	0.541	0.536	0.557	0.528
	0.5	10	0.653	0.480	0.584	0.587	0.624	0.428	0.676	0.530	0.550	0.568
		50	**0.574**	**0.605**	**0.735**	**0.462**	**0.630**	**0.447**	**0.465**	**0.538**	**0.611**	**0.563**
0.05	0.01	10	0.502	0.495	0.469	0.450	0.473	0.492	0.491	0.518	0.508	0.489
		50	0.477	0.488	0.487	0.471	0.479	0.522	0.479	0.504	0.504	0.490
	0.05	10	0.497	0.405	0.584	0.510	0.451	0.521	0.460	0.490	0.459	0.486
		50	0.462	0.493	0.439	0.456	0.535	0.491	0.503	0.501	0.493	0.486
	0.1	10	0.544	0.505	0.452	0.456	0.472	0.493	0.467	0.513	0.476	0.486
		50	0.544	0.445	0.521	0.447	0.565	0.534	0.508	0.498	0.507	0.508
	0.5	10	0.512	0.532	0.532	0.430	0.476	0.537	0.533	0.524	0.606	0.520
		50	0.498	0.465	0.488	0.531	0.525	0.532	0.556	0.582	0.554	0.526
0.1	0.01	10	0.469	0.530	0.500	0.443	0.498	0.457	0.501	0.478	0.482	0.484
		50	0.468	0.484	0.439	0.509	0.515	0.548	0.520	0.479	0.400	0.485
	0.05	10	0.492	0.479	0.476	0.463	0.506	0.463	0.467	0.508	0.551	0.489
		50	0.459	0.534	0.527	0.472	0.559	0.484	0.479	0.484	0.524	0.502
	0.1	10	0.406	0.488	0.448	0.514	0.436	0.511	0.489	0.524	0.558	0.486
		50	0.471	0.443	0.529	0.486	0.455	0.552	0.561	0.493	0.532	0.502
	0.5	10	0.515	0.336	0.454	0.486	0.514	0.526	0.518	0.555	0.563	0.496
		50	0.478	0.513	0.575	0.500	0.603	0.531	0.539	0.536	0.611	0.543
0.5	0.01	10	0.462	0.447	0.467	0.500	0.486	0.423	0.483	0.430	0.478	0.464
		50	0.532	0.424	0.491	0.461	0.386	0.466	0.478	0.488	0.488	0.468
	0.05	10	0.483	0.448	0.565	0.513	0.479	0.486	0.473	0.519	0.450	0.491
		50	0.516	0.484	0.471	0.438	0.431	0.499	0.479	0.483	0.480	0.476
	0.1	10	0.457	0.519	0.509	0.484	0.517	0.476	0.490	0.480	0.470	0.489
		50	0.486	0.498	0.443	0.448	0.494	0.463	0.473	0.502	0.508	0.479
	0.5	10	0.417	0.411	0.450	0.540	0.513	0.505	0.520	0.491	0.475	0.480
		50	0.452	0.454	0.541	0.590	0.545	0.521	0.572	0.537	0.534	0.527

Dapat dilihat pada Tabel 4.15, dari nilai rata-rata nilai koherensi tiap model paramater yang ditentukan. Pada analisis lebih lanjut perlu dilihat dari rata rata nilai koherensi tiap parameter pada grafik yang dapat dilihat pada Gambar 4.6



(a) Parameter α **(b) Parameter β**

Gambar 4.6 Rata-rata Nilai Koherensi Tiap ParameterMenggunakan *Term Frequency Inverse Document Frequency*

Dapat dilihat dari Gambar 4.6a rata-rata nilai koherensi tertinggi dengan parameter α pada jumlah iterasi 10 dan 50 kali mendapatkan hasil yang sama yakni pada parameter α sebesar 0.01. Sedangkan rata-rata nilai koherensi terendah pada parameter α berada pada nilai α sebesar 0.5 Dari Gambar 4.6b diketahui bahwa rata-rata nilai koherensi terbaik pada penggunaan parameter β pada jumlah iterasi 10 dan 50 kali yakni pada parameter β sebesar 0.5.Sementara rata-rata nilai koherensi terendah pada penggunaan parameter β berada pada nilai β sebesar 0.01.

Dari grafik pada Gambar 4.6 dapat disimpulkan bahwa semakin kecil nilai paramater α maka akan mendapatkan nilai koherensi semakin baik. Berbanding terbalik dengan nilai parameter β, semakin besar nilai parameter β maka nilai koherensi yang didapatkan semakin baik. Sementara jika nilai α semakin tinggi dan nilai β semakin rendah maka nilai koherensi dari model yang didapatkan akan semakin buruk.

Dari seluruh uji coba parameter dengan pembobotan kata TF-IDF yang telah dilakukan, didapatkan hasil terbaik pada model dengan nilai parameter α sebesar 0.01, β sebesar 0.5, dan jumlah iterasi sebanyak 50 kali pada jumlah topik

4 didapatkan nilai koherensi sebesar 0.735. Nilai tersebut merupakan hasil rata rata dari nilai koherensi tiap topik. Nilai koheresi tiap topik dapat dilihat pada Tabel 4.16

Tabel 4.16 Nilai Koherensi Tiap Topik Pembobotan *Term Frequency Inverse Document Frequency*

Id Topik	Nilai Koherensi
1	0.897
2	0.765
3	0.813
4	0.466

Dapat dilihat dari Tabel 4.15 nilai koherensi pada tiap topik cenderung berbeda. Topik dengan nilai koherensi terendah ada pada topik 4 dengan nilai koherensi sebesar 0.466. Sementara, nilai koherensi paling tinggi terdapat pada topik 1 dengan nilai koherensi sebesar 0.897. Artinya topik 1 menjadi topik yang terbaik.

4.7. Analisis pemodelan Topik

Setelah dilakukannya seluruh uji coba parameter mengunakan pembotan kata TF dan pembobotan kata TF-IDF. Didapatkan model terbaik menghasilkan 4 topik. Oleh karena itu, pembahasan selanjutnya akan berfokus dengan 4 topik yang telah terbentuk tersebut. Hasil model kata 4 topik dapat dilihat pada Tabel 4.17

Tabel 4.17 Model Kata Pada Tiap Topik

ID Topik	Model Topik
1	0.014*”malang” + 0.013*” kereta” + 0.012*“tiket”+0.013*”Informan” + 0.012*”booking” + 0.012*“kode” + 0.009*”periksa” + 0.008*”nyaman” + 0.008*”pergi”
2	0.014*”khusus” + 0.012*” tarif” + 0.012*“ekonomi” + 0.013*”ribu” + 0.011*”rute” + 0.011*“kereta” + 0.009*”eksekutif” + 0.008*”kelas” + 0.007*”mutiara”
3	0.014*”vaksin” + 0.013*” link” + 0.012*“wajib”+ 0.011*”lengkap” + 0.010*”syarat” + 0.010*“antigen” + 0.009*”tumpang” + 0.008*”kereta” + 0.008*”hasil”
4	0.015*”sehat” + 0.013*” jaga” + 0.012*“ketidaknyamannan”+ 0.013*”cek” + 0.012*”kendala” + 0.012*“kait” + 0.009*”tugas” + 0.007*”tindak” + 0.006*”alami”

Dapat dilihat pada Tabel 4.17,merupakan pemodelan kata pada tiap topik dengan 10 kata teratas. Tiap topik menghasilkan probabilitas kata kata yang cenderung berbeda. Namun, terdapat beberapa kata yang dapat masuk ke dalam beberapa topik, salah satu contohnya adalah ”kereta”. Dapat dilihat dari tabel 4.17 kata ”kereta” muncul pada topik 1 dan topik 3. Untuk memahami pemodelan topik yang terbentuk perlu dilakukannya analisis model topik yakni sebagai berikut:

4.7.1. Model Topik 1 LDA

Pada topik 1, dapat dilihat dari Tabel 4.17 terdapat kata ”kereta”, ”tiket”, ”booking” dan lain sebagainya. Dari kata kata tersebut, dapat disimpulkan bahwa topik 1 membahas seputar pemesanan tiket kereta. Kata - kata lainnya yang

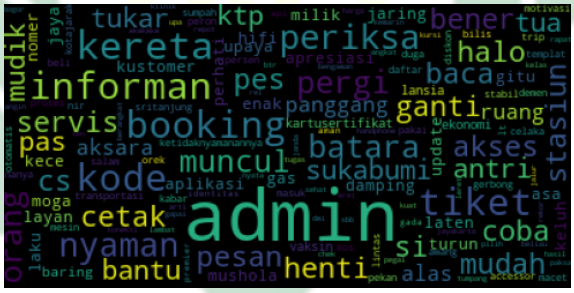
termasuk dalam topik 1 divisualkan menggunakan *word cloud*, ukuran kata yang tertulis di dalam *word cloud* menggunakan persamaan 2.20. Sebagai contoh, menghitung ukuran kata ”tiket” dengan variabel $f_{\max} = 50$, $t_{\max} = 444$, $t_{\min} = 5$, dan $t_{\text{tiket}} = 250$ sebagai berikut:

$$H_{\text{tiket}} = \frac{50\,(250 - 5)}{444 - 5}$$

$$= 12.25$$

$$= 12$$

Sehingga didapatkan hasil ukuran kata ”tiket” yakni sebesar 12. Untuk hasil visualisasi kata yang termasuk dalam topik 1 dapat dilihat pada Gambar 4.7.



Gambar 4.7 Visualisasi Kata Topik 1

Dapat dilihat dari Gambar **??** terdapat kata-kata yang mendukung topik 1 merupakan topik seputar pemesanan tiket kereta yakni ”cetak”, ”akses”, ”antri”, ”tukar”, dan lain sebagainya. PT Kereta Api Indonesia (Persero) memiliki dua akses untuk pembelian tiket yakno secara offline dengan membeli langsung ke loket, dan secara online pada aplikasi. Contoh tweet yang mengandung dominan topik 1 dengan probabilitas sebesar 0.65 dapat dilihat pada Tabel 4.18

Tabel 4.18 Contoh Tweet Mengandung Topik 1

Id	*Content*
1607913628061270000	Min @KAI121 untuk tiket potongan go show relasi SBY gubeng - purwokerto apakah ada?

4.7.2. Model Topik 2 LDA

Pada topik 2, dapat dilihat dari Tabel 4.17 terdapat kata ”tarif”, ”kelas”, ”ekonomi”, ”eksekutif” dan lain sebagainya. Dari kata kata tersebut, dapat disimpulkan bahwa topik 2 membahas seputar kelas pelayanan kereta. Kata - kata lainnya yang termasuk dalam topik 2 divisualkan menggunakan *word cloud*, ukuran kata yang tertulis di dalam *word cloud* menggunakan persamaan 2.20. Sebagai contoh, menghitung ukuran kata ”ekonomi” dengan variabel $f_{\max} = 50$, $t_{\max} = 444$, $t_{\min} = 5$, dan $t_{\text{ekonomi}} = 385$ sebagai berikut:

$$\begin{aligned} H_{\text{ekonomi}} &= \frac{50\,(385-5)}{444-5} \\ &= 43.28 \\ &= 43 \end{aligned}$$

Sehingga didapatkan hasil ukuran kata ”ekonomi” yakni sebesar 43. Untuk hasil visualisasi kata yang termasuk dalam topik 2 dapat dilihat pada Gambar 4.8.

Gambar 4.8 Visualisasi Kata Topik 2

Dapat dilihat dari Gambar 4.8 terdapat kata-kata yang mendukung topik 2 merupakan topik seputar kelas pelayanan kereta yakni ”premier”, ”bisnis”, ”kelas”, ”kursi”, dan lain sebagainya. Seperti yang diketahui PT Kereta api sendiri memiliki beberapa kelas pelayanan diantaranya adalah ekonomi, premium, bisnis, eksekutif, dan luxury. Contoh tweet yang mengandung dominan topik 2 dengan probabilitas sebesar 0.54 dapat dilihat pada Tabel 4.19

Tabel 4.19 Contoh Tweet Mengandung Topik 2

Id	*Content*
1605500936717090000	@ryan_risanto @ZakyMbarok @KAI121 Sdikit informasi dan koreksi tambahan saja, kalau KA singasari menggunakan Kereta Ekonomi yang dimodifikasi interiornya seperti kereta Premium

4.7.3. Model Topik 3 LDA

Pada topik 3, dapat dilihat dari Tabel 4.17 terdapat kata ”vaksin”, ”antigen”, ”syarat” dan lain sebagainya. Dari kata kata tersebut, dapat disimpulkan bahwa topik 3 membahas seputar protokol kesehatan. Kata - kata lainnya yang termasuk dalam topik 3 divisualkan menggunakan *word cloud*, ukuran kata yang tertulis di dalam *word cloud* menggunakan persamaan 2.20. Sebagai contoh, menghitung ukuran kata ”vaksin” dengan variabel $f_{\max} = 50$, $t_{\max} = 444$,

$t_{\min} = 5$, dan $t_{\text{vaksin}} = 322$ sebagai berikut:

$$H_{\text{vaksin}} = \frac{50\,(322 - 5)}{444 - 5}$$

$$= 36.10$$

$$= 36$$

Sehingga didapatkan hasil ukuran kata ”vaksin” yakni sebesar 36. Untuk hasil visualisasi kata yang termasuk dalam topik 3 dapat dilihat pada Gambar 4.9.

Gambar 4.9 Visualisasi Kata Topik 3

Dapat dilihat dari Gambar 4.9 terdapat kata-kata yang mendukung topik 3 merupakan topik seputar protokol kesehatan yakni ”vaksin”, ”dosis”, ”rapid”, ”antigen”, dan lain sebagainya. Seperti yang diketahui pada tahun 2022 Indonesia masih mengalami pandemi Covid-19 sehingga mengharuskan PT Kereta Api Indonesia (Persero) meningkatkan protokol kesehatan. Contoh tweet yang mengandung dominan topik 3 dengan probabilitas sebesar 0.778 dapat dilihat pada Tabel 4.20

Tabel 4.20 Contoh Tweet Mengandung Topik 3

Id	*Content*
1608492378725450000	@KAI121 min,mau tanya untuk naik kereta baru vaksin ke 2 sudah boleh atau masih butuh surat keterangan dokter?

4.7.4. Model Topik 4 LDA

Pada topik 4, dapat dilihat dari Tabel 4.17 terdapat kata ”tindak”, ”alami”, ”ketidaknyamanan” dan lain sebagainya. Dari kata kata tersebut, dapat disimpulkan bahwa topik 4 membahas seputar penyampaian keluhan. Kata - kata lainnya yang termasuk dalam topik 4 divisualkan menggunakan *word cloud*, ukuran kata yang tertulis di dalam *word cloud* menggunakan persamaan 2.20. Sebagai contoh, menghitung ukuran kata ”alami” dengan variabel $f_{\max} = 50$, $t_{\max} = 444$, $t_{\min} = 5$, dan $t_{\text{alami}} = 58$ sebagai berikut:

$$H_{\text{alami}} = \frac{50\,(58-5)}{444-5}$$
$$= 6,03$$
$$= 6$$

Sehingga didapatkan hasil ukuran kata ”alami” yakni sebesar 6. Untuk hasil visualisasi kata yang termasuk dalam topik 4 dapat dilihat pada Gambar 4.10.

Gambar 4.10 Visualisasi Kata Topik 4

Dapat dilihat dari Gambar 4.10 terdapat kata-kata yang mendukung topik 4 merupakan topik seputar penyampaian keluhan yakni ”kendala”, ”lapor”, ”kondektur”, ”tugas”, dan lain sebagainya. PT Kereta Api Indonesia (Persero) memiliki berbagai macam cara untuk menyampaikan keluhan masyarakat, sebagai contoh adalah PT Kereta Api Indonesia (Persero) menaruh nomor telepon kondektur di depan dan di belakang setiap gerbong kereta. Contoh tweet yang mengandung dominan topik 4 dengan probabilitas sebesar 0.851 dapat dilihat pada Tabel 4.21

Tabel 4.21 Contoh Tweet Mengandung Topik 4

Id	*Content*
1607737635795270000	Kenapa sih beli tiket kereta di KAI Access selalu crash aplikasinya. Dibuka di hp sendiri crash tapi dibuka di hp temen lancar. Ini yang salah hpnya atau aplikasinya ya? soalnya udah beberapa kali install ulang tetep aja crash mulu terutama pas mau masukin data penumpang @KAI121

4.8. Integrasi Keislaman

Allah menciptakan dunia beserta keanekaragaman isinya. Oleh karena itu kita sebagai manusia perlu mensyukuri kenikamatan tersebut dengan menjelajahi

bagian yang ada didunia. Di dalam Al-Qur'an terdapat perintah menjelajahi isi dunia tertulis dalam surah Al-Mulk ayat 15 yang berbunyi:

هُوَ ٱلَّذِى جَعَلَ لَكُمُ ٱلْأَرْضَ ذَلُولًا فَٱمْشُوا۟ فِى مَنَاكِبِهَا وَكُلُوا۟ مِن رِّزْقِهِۦ ۖ وَإِلَيْهِ ٱلنُّشُورُ ﴿١٥﴾

artinya: Dialah yang menjadikan bumi untuk kamu yang mudah di jelajahi, maka jelajahi lah di segala penjurunya dan makanlah sebagaian dari rezeki-Nya. Dan hanya kepada-Nya lah kamu (kembali setelah) dibangkitkan (QS. Al-Mulk: 15).

Isi dari surah AL-Mulk ayat 15 menandakan bahwa, Allah telah memberikan manusia rezeki yang tersebar di penjuru dunia. Oleh karena itu, kita sebagai manusia sepatutnya berikhtiar untuk mendapatkan dan menyukuri rezeki yang telah Allah berikan. Dalam menjelajahi dunia, manusia perlu adanya kemajuan dalam bidang transportasi. Salah satu transportasi yang memiliki kemajuan teknologi yang pesat dan digemari masyarakat adalah kereta api.

Kereta api di Indonesia yang dikelola oleh PT Kereta Api Indonesia (Persero) menjadi wadah bagi pengguna jasa nya sebagai penghubung jarak dalam mecari rezeki. Oleh karena itu, PT Kereta Api Indonesia hendaknya selalu meningkatkan kualitas pelayanan. Dalam memberikan pelayanan seharusnya PT Kereta Api Indonesia (Persero) selalu memberikan yang terbaik. Pelayanan yang baik merupakan tindakan melayani dengan hati yang iklash dan penuh dengan kesenangan. Salah satu memberikan pelayanan yang baik yakni melayani dengan senyuman. Di dalam islam memberikan senyuman merupakan bagian dari sedekah, hal tersebut tercantum dalam hadist sebagai berikut:

عَنْ أَبِي ذَرٍّ قَالَ قَالَ رَسُولُ اللَّهِ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ تَبَسُّمُكَ فِي وَجْهِ أَخِيكَ لَكَ صَدَقَةٌ ، وَأَمْرُكَ بِالْمَعْرُوفِ وَنَهْيُكَ عَنْ الْمُنْكَرِ صَدَقَةٌ ، وَإِرْشَادُكَ الرَّجُلَ فِي أَرْضِ الضَّلَالِ لَكَ صَدَقَةٌ ، وَبَصَرُكَ لِلرَّجُلِ الرَّدِيءِ الْبَصَرِ لَكَ صَدَقَةٌ ، وَإِمَاطَتُكَ الْحَجَرَ وَالشَّوْكَةَ وَالْعَظْمَ عَنْ الطَّرِيقِ لَكَ صَدَقَةٌ ، وَإِفْرَاغُكَ مِنْ دَلْوِكَ فِي دَلْوِ أَخِيكَ لَكَ صَدَقَةٌ .

artinya: dari Abu Dzarr ia berkata ”Rasulullah shallallahu ’alaihi wasallam bersabda: Senyummu kepada saudaramu merupakan sedekah, engkau berbuat ma’ruf dan melarang dari kemungkaran juga sedekah, engkau menunjukkan jalan kepada orang yang tersesat juga sedekah, engkau menuntun orang yang berpenglihatan kabur juga sedekah, menyingkirkan batu, duri dan tulang dari jalan merupakan sedekah, dan engkau menuangkan air dari embermu ke ember saudaramu juga sedekah” (HR. Tirmidzi no 1879).

Isi dari hadist riwayat Tirmidzi no 1879 menunjukkan keutamaan tersenyum dan menampakkan muka manis kepada orang lain. Memberikan senyuman kepada orang lain merupakan sedekah. Seperti halnya dalam melakukan pelayanan hendaknya menerapkan apa yang digambarkan pada hadist tersebut. Jika dalam melayani pelanggan dengan senyuman, pelanggan akan merasa senang dan nyaman. Oleh karena itu dalam memberikan pelayanan seharusny memberikan yang terbaik atau berkualitas. Jangan memberikan pelayanan berupa pelayan yang buruk atau tidak berkualitas.

Namun, usaha memberikan usaha yang maksimal belum tentu memberikan kepuasaan untuk setiap pelanggan PT Kereta Api Indosenia. Karena, pada dasarnya manusia memiliki sifat yang tidak pernah puas seperti apa yang disampaikan pada

hadist berikut:

Dari Abu Hurairah, Rasulullah shallallahu'alaihi wa sallam bersabda,

لَوْ كَانَ لِابْنِ آدَمَ وَادِيَانِ مِنْ مَالٍ لَابْتَغَى ثَالِثًا ، وَلاَ يَمْلأُ جَوْفَ ابْنِ آدَمَ إِلاَّ التُّرَابُ ، وَيَتُوبُ اللَّهُ عَلَى مَنْ تَابَ

artinya: Seandainya manusia diberi dua lembah berisi harta, tentu ia masih menginginkan kembali lembah yang ketiga. Yang bisa memenuhi dalam perut manusia hanyalah tanah. Allah tentu akan menerima taubat bagi siapa saja yang ingin bertaubat (HR. Bukhari n0 1048)

Dalam hadist riwayat Bukhari no 1048 dijelaskan bahwa manusia memiliki rasa ambisi yang terus menerus dan tidak pernah puas. Sifat manusia yang tidak pernah puas menyebabkan adanya opini terhadap suatu objek salah satunya terhadap PT Kereta Api Indonesia (Persero). Walaupun kemajuan teknologi transportasi kereta api di Indonesia terbilang cukup pesat. Masih ada saja opini opini publik terhadapnya, baik opini baik maupun opini buruk.

Sebagai pengguna jasa yang memiliki opini wajar saja bila disampaikan. Karena agama islam pun mengajarkan menyampaikan opini untuk menyelesaikan masalah. Tetapi, dalam menyampaikan opini perlu kaidah kaidah tertentu. Salah satunya, adalah menyampaikan opini sebaiknya berdasarkan pada kebenaran dan tidak dipengaruhi oleh hawa nafsu. Kaidah tersebut menggambarkan seseorang yang beropini memiliki pemikiran yang kritis dan tidak sembarangan untuk berbicara. Sehingga opini yang telah disampaikan dapat ditindaklanjuti oleh PT Kereta Api Indonesia (Persero).

Allah SWT melarang hambanya untuk membuat opini yang tidak benar atau tidak valid. Sebagaimana firman Allah tentang larangan menyebarkan opini atau

berita yang tidak benar yang tertuang dalam surah Al-Hujarat ayat 6 yang berbunyi:

يَٰٓأَيُّهَا ٱلَّذِينَ ءَامَنُوٓا۟ إِن جَآءَكُمْ فَاسِقٌۢ بِنَبَإٍ فَتَبَيَّنُوٓا۟ أَن تُصِيبُوا۟ قَوْمًۢا بِجَهَٰلَةٍ فَتُصْبِحُوا۟ عَلَىٰ مَا فَعَلْتُمْ نَٰدِمِينَ

artinya: Wahai orang-orang yang beriman! Jika seseorang yang fasik datang kepadamu membawa suatu berita, maka telitilah kebenarannya, agar kamu tidak mencelakakan suatu kaum karena kebodohan (kecerobohan), yang akhirnya kamu menyesali perbuatanmu itu (QS. Al-Hujarat: 6).

Surah Al-Hujarat ayat 6 menjelaskan bahwa kita sebagai manusia yang beriman untuk berhati-hati dan berpikir secara bijak ketika menerima maupun menyampaikan berita dari atau kepada seseorang yang tidak dapat dipercaya. Ayat ini menekankan bahwa seseorang tidak boleh dengan mudahnya menyalahkan atau menyebabkan kemudaratan kepada suatu kelompok atau masyarakat hanya berdasarkan berita yang belum terverifikasi secara akurat atau dapat disimpulkan sebagai opini pribadi yang tidak jelas asal usulnya. Begitu pula sebaliknya, orang-orang beriman diharapkan untuk mencari kebenaran, memahami konteks, dan memastikan keabsahan berita sebelum mengambil tindakan atau memberikan respons terhadapnya. Oleh karena itu, kita sebagai masyarakat harus bijak dalam menanggapi atau menyampaikan opini di dalam media sosial. Tidak terkecuali tentang menyampaikan pendapat pada PT Kereta Api Indonesia (Persero) di dalam media sosial twitter.

BAB V

PENUTUP

5.1. Kesimpulan

Dari penelitian yang telah dilakukan dapat disimpulkan bahwa twitter merupakan salah satu media penghubung antara pengguna layanan dengan perusahaan PT Kereta Api Indonesia (Persero). Hal tersebut dibuktikkan dari banyak data yanhg di dapatkan dalam kurun waktu 01 Januari 2022 hingga 31 Desember 2022, dan beberapa sampel tweet yang berisikan keluhan dan pertanyaan. Dari hasil penelitian yang dilakukan dapat disimpulkan lagi menjadi beberapa poin sebagai berikut:

1. Pada penelitian ini dilakukan 2 pembobotan yang berbeda yakni *term frequency* dan *term frequency inverse document frequency*, dari hasil penelitian yang telah dilakukan pembobotan dengan *term frequency inverse document frequency* mendapatkan nilai koherensi lebih baik dengan 4 topik yakni sebesar 0.735 menggunakan parameter α sebesar 0.01, β sebesar 0.5, dan jumlah iterasi sebanyak 50 kali . Sementara pada pembobotan *term frequency* nilai koherensi terbesar hanya 0.69 dengan jumlah topik sebanyak 10 parameter α sebesar 0.05, β sebesar 0.01, dan jumlah iterasi sebanyak 50 kali.

2. Dari 4 topik terbaik dengan pembobotan *term frequency inverse document frequency* didapatkan nilai koherensi tiap topiknya yakni topik 1 sebesar 0.897, topik 2 sebesar 0.765, topik 3 sebesar 0,813 dan topik 4 sebesar

0.466.

3. Hasil dari pengelompokan topik opini masyarakat terhadap PT Kereta Api Indonesia (Persero) menggunakan metode LDA menghasil kan 4 pembahasan topik yakni topik 1 membahas tentang pemesanan tiket kereta kereta, topik 2 membahas tentang kelas pelayanan, topik 3 membahas seputar protokol kesehatan, dan topik 4 membahas seputar penyampaian keluhan.

5.2. Saran

Dalam penelitian ini, penulis menyadari memiliki banyak keterbatasan dan kekurangan pada penelitian yang telah penulis selesaikan. Oleh karena itu, berdasarkan penelitian ini, penulis menyarankan untuk melakuan pengembangan penelitian ini agar menjadi lebih baik dengan beberapa poin sebagai berikut:

1. Pada penelitian ini, data diambil dari media sosial twitter. Diharapkan dalam penelitian berikutnya dapat mengambil data dari media sosial lain seperti instagram atau facebook.

2. Pada penelitian ini, data yang digunakan hanya tweet dengan kata pencarian ”@KAI121”. Pada penelitian berikutnya dapat menggunakan data dari akun resmi yang lainnya seperti ”@commuterline”, ”@keretaapikita” dan lain sebagainya.

3. Pada penelitian ini mendapatkan nilai koherensi terbaik sebesar 0.735. Diharapkan pada penelitian selanjutnya dapat menggunakan metode pembobotan kata lainnya untuk mendapatkan nilai koherensi yang lebih baik.

4. Dapat mengembangkan pemodelan topik dengan metode lainnya.

DAFTAR PUSTAKA

Acharya, A. S., Prakash, A., Saxena, P., and Nigam, A. (2013). Sampling: Why and How of I t? *Indian Journal of Medical Specialities*, 4(2):3–7.

Alash, H. M. and Al-Sultany, G. A. (2020). Improve topic modeling algorithms based on Twitter hashtags. *Journal of Physics: Conference Series*, 1660(1).

Albattah, W. (2016). The Role of Sampling in Big Data A nalysis. *ACM International Conference Proceeding Series*, pages 1–5.

Aletras, N. and Stevenson, M. (2013). Evaluating Topic Coherence Using Distributional Semantics. *Proceedings of the 10th International Conference on Computational Semantics, IWCS 2013 - Long Papers*, (2009).

Blei, D. M. (2012). Introduction to Probabilstic Topic Models. *Journal of Machine Learning*, 64(3):277–278.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data. *Journal of Machine Learning Research 3*, 3:139–159.

Brito, M., Cunha, J., and Saraiva, J. (2021). Identification of microservices from monolithic applications through topic modelling. *Proceedings of the ACM Symposium on Applied Computing*, pages 1409–1418.

Brzustewicz, P. and Singh, A. (2021). Sustainable Consumption in Consumer Behavior in The Time of COVID-19: Topic Modeling on Twitter Data Using LDA. *Energies*, 14(18).

BUMN (2022). Kementerian Badan Usaha Milik Negara. https://bumn.go.id/post/rayakan-internalisasi-budaya-perusahaan-kai-gelar.

CNN Indonesia (2017). Kemajuan Perkeretaapian RI Buah Konsistensi Inovasi. https://www.cnnindonesia.com/nasional/20190927111517-293-434563/kemajuan-perkeretaapian-ri-buah-konsistensi-inovasi.

Curiskis, S. A., Drake, B., Osborn, T. R., and Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing and Management*, 57(2).

Darling, W. M. (2011). A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1–10.

Darussalam and Arief, G. (2017). Topic Modeling for Support Ticket using Latent Dirichlet Allocation. *Resti*, 1(1):19–25.

El Rahman, S. A., Alotaibi, F. A., and Alshehri, W. A. (2019). Sentiment Analysis of Twitter Data. *2019 International Conference on Computer and Information Sciences, ICCIS 2019*.

Erman, L. M. and Sitanggang, I. S. (2016). Clustering Undergraduate Computer Science Student Final Project Based on Frequent Itemset. *International Journal of Information Technology and Computer Science*, 8(11):1–7.

Fida, I. (2022). Pelcehan Seksual oleh Staff PT Kereta Api Indonesia di Stasiun Ciamis.

Fikriyah, S. N. and Sibaroni, Y. (2022). Identify User Behavior based on Tweet Type on twitter Platform using Mean Shift Clustering. *Media Informatika Budidarma*, 6:1396–1403.

Henry, G. (2016). Sample Selection Approaches. *Practical Sampling*, pages 17–32.

Indonesia, P. K. A. (2022). KAI Akan Remajakan Kereta Ekonomi untuk Tingkatkan Kenyamanan.

Info, C. (2022). Tegas Sikapi Dugaan Pelecehan Seksual di Stasiun Ciamis, PT KAI Bebastugaskan Oknum Petugas Cleaning Service.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2018). Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a survey. *Multimedia Tools and Applications*, 78:183–198.

Jin, Y. (2017). Development of Word Cloud Generator Software Based on Python. *Procedia Engineering*, 174:788–792.

Jo, T. (2019). *Text Mining Concepts, Implementation, and Big Data Challenge*. Poland, 45 edition.

Kalepalli, Y., Tasneem, S., Teja, P. D. P., and Manne, S. (2020). Effective Comparison of LDA with LSA for Topic Modelling. *Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020*, (Iciccs):1245–1250.

Kang, J., Lee, J., Jang, D., and Park, S. (2019). A Methodology of Partner Selection for Sustainable Industry-University Cooperation Based on LDA Topic Model. *Sustainability (Switzerland)*, 11(12):1–16.

Kaur, A. and Chopra, D. Comparison of Text Mining Tools. *2016 5th International Conference on Reliability, Infocom Technologies and Optimization, ICRITO 2016: Trends and Future Directions*, pages 186–192.

Kearney, M. (2019). Rtweet: Collecting and Analyzing Twitter Data. *Journal of Open Source Software*, 4(42):1829.

Kherwa, P. and Bansal, P. (2020). Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems*, 7(24):1–16.

Kumar, L. and Bhatia, P. K. (2013). Text Mining: Concepts, Process, and Applications. 4(3):36–39.

Kwak, H., Park, H. M., Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a Social Network or a News Media. 4.

Maindonald, J. (2007). *Pattern Recognition and Machine Learning*, volume 17.

Mohammadi, E. and Karami, A. (2022). Exploring Research Trends in Big Data Across Disciplines: A Text Mining Analysis. *Journal of Information Science*, 48(1):44–56.

Mohammed, S. H. and Al-Augby, S. (2020). LSA & LDA Topic Modeling Classification: Comparison Study on E-Books. *Indonesian Journal of Electrical Engineering and Computer Science*, 19(1):353–362.

Muhammad, A. M., Prawiradiredja, S., and Fitriyah, I. (2018). Corporate Value: Persona pada Company Profile PT. Kereta Api Indonesia. *Jurnal Komunikasi Profesional*, 2(1):29–37.

Mutia, A. (2022). BPS: Jumlah Penumpang Kereta Api Turun 5,64 Persen di Agustus 2022.

Natalia, C., Suprata, F., Surbakti, F. P. S., and Clarence, S. (2021). Penentuan Standar Spesifikasi Kerja di Café Berdasarkan Big Data dengan Metode LDA dan AHP. *Jurnal Rekayasa Sistem Industri*, 10(2):211–226.

Negara, E. S. and Triadi, D. (2021). Topic Modeling Using Latent Dirichlet Allocation (LDA) on Twitter Data with Indonesia Keyword. *Bulletin of Social Informatics Theory and Application*, 5(2):124–132.

Nimastiti, F. and Julianto, E. (2022). Indonesian Twitter Stopwords. https://github.com/Braincore-id/IndoTWEEST/blob/main/stopwords_twitter.csv.

Noorca, D. (2022). KAI Akan Remajakan Unitnya Usai Kursi Berhadapan Kereta Ekonomi Viral Dikeluhkan Warganet.

Nugroho, S. A., Bachtiar, F. A., and Wihandika, R. C. (2021). 53 Aspect Extraction in E-Commerce Using Latent Dirichlet Allocation (Lda) With Term Frequency-Inverse Document Frequency (Tf-Idf). *Jurnal Ilmiah KURSOR*, 11(2):53–62.

Onan, A. and Tocoglu, M. A. (2021). Model and Stacked Bidirectional LSTM Based Framework for Sarcasm Identification. pages 1–23.

Owen, L. (2020a). Indonesian Stopwords. https://github.com/louisowen6/NLP_bahasa_resources/blob/master/combined_stop_words.txt.

Owen, L. (2020b). Kata Dasar Bahasa Indonesia. https://github.com/louisowen6/NLP_bahasa_resources/blob/master/combined_root_words.txt.

Owen, L. (2020c). Kata Gaul Bahasa Indonesia. https://github.com/louisowen6/NLP_bahasa_resources/blob/master/combined_slang_words.txt.

Papanikolaou, Y., Foulds, J. R., Rubin, T. N., and Tsoumakas, G. (2017). Dense Distributions from Sparse Samples: Improved Gibbs Sampling Parameter Estimators for LDA. *Journal of Machine Learning Research*, 18:1–58.

Pratama, M. O., Satyawan, W., Jannati, R., Pamungkas, B., Raspiani, Syahputra, M. E., and Neforawati, I. (2019). The Sentiment Analysis of Indonesia Commuter Line Using Machine Learning Based on Twitter Data. *Journal of Physics: Conference Series*, 1193(1).

Pratiwi, I. Y. R. (2022). Hoax News Identification Using Machine Learning Model From Online Media in Bahasa Indonesia. *MATRIX : Jurnal Manajemen Teknologi dan Informatika*, 12(2):58–67.

Ranjbari, M., Saidani, M., Shams Esfandabadi, Z., Peng, W., Lam, S. S., Aghbashlo, M., Quatraro, F., and Tabatabaei, M. (2021). Two decades of research on waste management in the circular economy: Insights from bibliometric, text mining, and content analyses. *Journal of Cleaner Production*, 314(March):128009.

Reisenbichler, M. and Reutterer, T. (2019). Topic modeling in marketing: recent advances and research opportunities. *Journal of Business Economics*, 89(3):327–356.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pages 399–408.

Rohman, H. N. and Asror, I. (2019). Automatic Detection of Argument Components in Text Using Multinomial Nave Bayes Clasiffier Automatic

Detection of Argument Components in Text Using Multinomial Nave Bayes Clasiffier.

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., and Both, A. (2014). Evaluating Topic Coherence Measures. pages 1–4.

Sahria, Y. and Fudholi, D. H. (2017). Analisis Topik Penelitian Kesehatan di Indonesia Menggunakan Metode Topic Modeling LDA (Latent Dirichlet Allocation). *Jurnal Rekayasa Sistem dan Teknologi Informasi*, 1(3):336–344.

Saif, H., Fernandez, M., He, Y., and Alani, H. (2014). On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 810–817.

Samant, S. S., Bhanu Murthy, N. L., and Malapati, A. (2019). Improving Term Weighting Schemes for Short Text Classification in Vector Space Model. *IEEE Access*, 7:166578–166592.

Sarioglu, E., Yadav, K., and Choi, H.-A. (2013). Topic Modeling Based Classification of Clinical Reports. *Proceedings of the ACL Student Research Workshop*, pages 67–73.

Schmidt, J.-H. (2014). Twitter and the Rise of Personal Publics. *Twitter and sochiety*, pages 3–15.

Setiabudi, R., Iswari, N. M. S., and Rusli, A. (2021). Enhancing Text Classification Performance by Preprocessing Misspelled Words in Indonesian language. *Telkomnika (Telecommunication Computing Electronics and Control)*, 19(4):1234–1241.

Steyvers, M. and Griffiths, T. (2010). Probalistic Topic Models. *Latent Semantic Analysis: A Road To Meaning*, 3(3):993–1022.

Thakur, K. and Kumar, V. (2022). Application of Text Mining Techniques on Scholarly Research Articles: Methods and Tools. *New Review of Academic Librarianship*, 28(3):279–302.

Thakur, N. (2022). Twitter Big Data as a Resource for Exoskeleton Research: A Large-Scale Dataset of about 140,000 Tweets from 2017–2022 and 100 Research Questions. *Analytics*, 1(2):72–97.

Trimastuti, W. (2017). An Analysis of Slang Words Used in Social Media. *Jurnal Dimensi Pendidikan dan Pembelajaran*, 5(2):64–68.

Twitter (2022). Twitter Analytics @KAI121.

Weidner, K., Lowman, J., Fleischer, A., Kosik, K., Goodbread, P., Chen, B., and Kavuluru, R. (2021). Twitter, Telepractice, and the COVID-19 Pandemic: A Social Media Content Analysis. *American Journal of Speech-Language Pathology*, 30(6):2561–2571.

Wolter, K. M. (1984). An Investigation of Some Estimators of Variance for Systematic Sampling. *Journal of the American Statistical Association*, 79(388):781–790.