Indonesian Journal of Applied Technology, Vol: 1, No 2, 2024, Page: 1-24

Fitur Ekstraksi pada Pemodelan Topik Menggunakan Metode Latent Dirichlet Allocation pada Peristiwa Kebocoran Data

Achmad Ariansyah, Uce Indahyanti*
Universitas Muhammadiyah Sidoarjo

Abstrak: Penelitian ini bertujuan untuk mencari fitur ekstraksi terbaik serta menerapkan pemodelan topik dari data Twitter tentang kebocoran data pribadi, salah satu trending topik akibat ulah oknum hacker bjorka dimana data yang tersebar merupakan data penting seperti NIK dan SIM Card rakyat Indonesia. Penelitian dilakukan dengan metode Latent Dirichlet Allocation (LDA) menggunakan fitur ekstraksi Bag of Word (BoW) dan TF-IDF, dan data yang digunakan terdiri dari 11.067 tweet dari platform twitter. Pemodelan dengan menggunakan fitur ekstraksi BoW menghasilkan score coherences terbaik bernilai 0.47 dengan 3 topik utama terkait kebocoran data seperti kominfo lindungi data pribadi, johnny g plate bertanggung jawab atas kasus kebocoran data ulah hacker bjorka dan perlindungan data pribadi rakyat memalui ruu pdp. Sementara itu, dengan fitur ekstraksi TF-IDF mendapatkan score coherences terbaik bernilai 0.47 dengan 5 topik utama, akan tetapi topik tersebut tidak dapat di interpretasikan dengan baik seperti menggunakan fitur ekstraksi BoW.

Keywords: Latent Dirichlet Allocation, Bag of Word, TF-IDF, Kebocoran Data

DOI: https://doi.org/10.47134/ijat.v1i2.3041
*Correspondence: Uce Indahyanti
Email: uceindahyanti@umsida.ac.id

Received: 14-04-2024
Accepted: 16-04-2024
Published: 27-04-2024

Abstract: *This research aims to find the best extraction features and apply topic modeling from Twitter data regarding personal data leaks, one of the trending topics due to the actions of hacker Bjorka where the data that is spread is important data such as the NIK and SIM cards of the Indonesian people. The research was carried out using the Latent Dirichlet Allocation (LDA) method using the Bag of Word (BoW) and TF-IDF extraction features, and the data used consisted of 11,067 tweets from the Twitter platform. Modeling using the BoW extraction feature produces the best coherence score of 0.47 with 3 main topics related to data leaks such as Kominfo protecting personal data, Johnny G Plate being responsible for the data leak case caused by hacker Bjorka and protecting people's personal data through the PDP bill. Meanwhile, with the TF-IDF extraction feature, the best coherence score was 0.47 with 5 main topics, but these topics could not be interpreted as well as using the BoW extraction feature.*

Keywords: *Latent Dirichlet Allocation, Bag of Word, TF-IDF, Data Lea*

Pendahuluan

Dalam beberapa tahun terakhir, ada satu tren yang muncul di masyarakat karena meningkatnya penggunaan teknologi informasi dan komunikasi yakni menyampaikan opini dan aspirasi melalui media online seperti situs, forum online, blog, dan jejaring sosial seperti Twitter(Kozlowski et al., 2021). Twitter merupakan media sosial populer yang

pertama kali dipublikasikan pada tahun 2006. User Twitter di Indonesia selalu bertambah di tiap tahun, bahkan Indonesia menjadi salah satu negara dengan jumlah pengguna Twitter terbesar di dunia. Menurut sumber We Are Social, jumlah pengguna Twitter di Indonesia mencapai 18,45 juta pada tahun 2022(John et al., 2022) .

Media sosial Twitter semakin populer karena banyak informasi yang dapat diakses dari sana. Pengguna menggunakannya untuk berbagai kebutuhan, termasuk kebutuhan publik, pemerintah, dan bisnis. Tweet yang dibuat oleh pengguna Twitter mencakup berbagai topik. Dari tweet-tweet tersebut, dapat ditemukan topik utama yang sedang banyak diperbincangkan oleh pengguna saat itu dengan melakukan analisis topik (Xue et al., 2020).

Peristiwa kebocoran data yang terjadi pada bulan September 2022 merupakan salah satu topik yang banyak dibicarakan di Twitter. Peristiwa ini merugikan masyarakat karena data-data privasi, seperti data pelanggan indihome, data registrasi sim card, dan data KPU RI yang berisi data penduduk sebanyak 105.003.428, terungkap ke publik. Data tersebut mencakup informasi seperti NIK, KK, nama lengkap, tempat dan tanggal lahir, jenis kelamin, alamat, dan usia . Hal ini menimbulkan rasa takut dalam masyarakat dan memancing pengguna aplikasi Twitter untuk menuliskan tweet tentang keresahan mereka, ketidakpercayaan terhadap pemerintah dalam melindungi data pribadi, dan berbagai topik lainnya. Peristiwa ini menjadi trending topik di Twitter pada bulan September, sehingga penulis tertarik untuk menjadikannya sebagai studi kasus dalam penelitian(Kang et al., 2019) .

Beberapa penelitian sejenis sebelumnya telah dilakukan, seperti penelitian yang dilakukan oleh Wirasakti, Permadi, Hartanto, dan Hartatik (2020) tentang membuat kata kunci otomatis di dalam artikel atau paper dengan menggunakan pemodelan topik. Tujuan penelitian tersebut adalah menemukan kata kunci yang cocok untuk digunakan dalam sebuah publikasi artikel dalam sebuah blog dengan menggunakan model LDA yaitu probabilitas yang dapat menghasilkan beberapa topik. Penelitian lain yang dilakukan oleh Guntur Budi Herwanto (2018) mengenai dokumen clustering dengan LDA dan ward hierarichal clustering (Tran et al., 2019).

Dari beberapa penelitian terdahulu tidak ada penelitian sejenis sebelumnya yang membahas topik peristiwa kebocoran data yang terjadi di Indonesia pada bulan September 2022. Akibat peristiwa tersebut, banyak sekali cuitan atau tweet dari masyarakat yang membahas beberapa topik yang berkaitan dengan kebocoran data tersebut. Oleh karena itu, dalam penelitian ini diusulkan untuk melakukan pemodelan topik pengguna Twitter mengenai peristiwa kebocoran data tersebut dengan menggunakan metode Latent Dirichlet Allocation. Tujuannya adalah untuk menentukan topic modeling dengan menggunakan topic coherence sebagai validasi topik yang dihasilkan(Han, 2020) .

Penelitian Terdahulu Dan Landasan Teori

2.1 Penelitian Terdahulu

Perbandingan beberapa penelitian sejenis terdahulu dapat dilihat pada tabel 2.1. berikut :

Tabel 1 *Perbandingan Penelitian Terdahulu*

No	Judul	Metode	Hasil
1.	Document Clustering Dengan Latent Dirichlet Allocation dan Ward Hierarichal Clustering (Guntur dkk, 2018)	Menggunakan metode Latent Dirichlet Allocation dan Ward Hierarichal Clustering	Kombinasi dari metode LDA dan Ward Hierarichal Clustering digunakan untuk mewakili vektor dokumen sebagai distribusi topik dengan hasil silhouette coefficient yang baik, yaitu 0.7.
2.	Pembuatan Kata Kunci Otomatis Dalam Artikel Dengan Pemodelan Topik (Wirasakti, 2020)	Menggunakan metode Latent Dirichlet Allocation	Memperoleh 4 topik yang memiliki nilai probabilitas tinggi dengan kata mesin, maksimal, varian, cx-8, mobil dan mazda dari data teks pada artikel sebuah website atau blog.
3.	Latent Dirichlet Allocation (LDA) Untuk Mengetahui Topik Pembicaraan Warganet Twitter Tentang Omnibus Law (Luvian dkk, 2021)	Menggunakan metode Latent Dirichlet Allocation	Subjek Omnibus Law memilki 5 topik dengan coherence score sebesar 0.5644

Penelitian yang akan dilakukan oleh penulis menggunakan metode yang sama yaitu LDA mengenai topik Peristiwa Kebocoran Data pada media social Twitter

2.2 Landasan Teori

2.2.1 Data

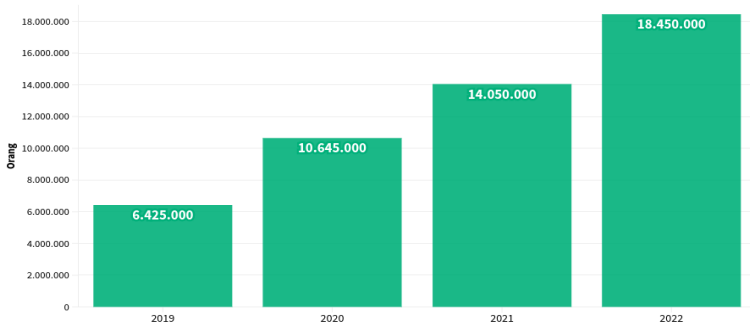
Data adalah kumpulan informasi atau fakta yang terdiri dari kata, frase, simbol, angka, dan lain-lain. Data diperoleh dari proses pencarian dan pengamatan yang tepat berdasarkan sumber tertentu. Bentuk lain dari pemahaman data adalah kumpulan informasi atau deskripsi dasar yang berasal dari suatu objek atau peristiwa. Menurut pendapat lain, data adalah rekaman fakta, konsep, atau instruksi yang perlu diolah untuk menghasilkan informasi yang dapat dipahami oleh orang. Data dapat disimpan dalam berbagai bentuk media komputer seperti video, gambar, audio, dan teks. Oleh karena itu, pengertian data pada era ini dapat

diperluas menjadi data berupa fakta, konsep, petunjuk, grafik, audio, dan video(Wang et al., 2020) .

Data pribadi juga merupakan suatu kumpulan data yang berisi informasi penting yang melekat pada seseorang. Data pribadi harus dilindungi karena merupakan hak privasi setiap orang. Hak privasi adalah hak konstitusional yang diatur dalam Undang-Undang Dasar Negara Republik Indonesia Tahun 1945. Hak konstitusional adalah kewajiban negara terhadap warga negaranya .

2.2.2 Twitter

Twitter merupakan sosial media yang menjadi situs berbagi informasi di Indoneisa. Selain itu, twitter juga menjadi media berkomunikasi beropini atau mengemukakan pendapat secara cepat. Kecepatan dan kemudahan twitter ini menjadikan medium pilihan bagi pengguna untuk berkomunikasi setiap hari. Twitter merupakan salah satu media sosial populer di dunia sejak pertama kali dipublikasikan pada tahun 2006 . Pengguna Twitter di Indonesia setiap tahunnya selalu bertambah, bahkan Indonesia menjadi salah satu negara dengan pengguna Twitter terbesar di dunia. Berdasarkan sumber We Are Social, jumlah pengguna Twitter di Indonesia mencapai 18,45 juta pada tahun 2022.



Gambar 1. *Grafik Pengguna Twitter*

2.2.3 Text Mining

Text mining adalah suatu metode analisis data yang menggunakan bahasa alami dengan teknik dan alat tertentu untuk merancang, menemukan, dan mengekstrak pengetahuan dari data yang tidak terstruktur. Dengan menggunakan penambangan teks, data tidak terstruktur dapat diolah menjadi data yang lebih terstruktur dan mudah dianalisis. Penambangan teks ini dapat mengubah kata atau kalimat untuk mempermudah analisis (Janmaijaya et al., 2021).Text mining dengan pencarian otomatis sangat berkaitan karena tujuan text mining dan pencarian otomatis yaitu untuk mendapatkan atau menghasilkan informasi yang berguna dari beberapa dokumen.

Text mining dapat memberikan solusi dari permasalahan seperti pemrosesan, pengorganisasian atau pengelompokkan dan menganalisa unstructured text dalam jumlah besar. Dalam memberikan solusi, text mining mengadopsi dan mengembangkan banyak teknik dari bidal lain, seperti Data mining, Information Retrieval, Statistik dan Matematik, Machine Learning, Linguistic, Natural Language Processing (NLP), dan Visualization. Tujuan dari text mining adalah untuk mendapatkan proses knowledge discovery pada koleksi dokumen yang besar untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Adapun tugas khusus dari text mining antara lain yaitu pengkategorisasian teks (text categorization) dan pengelompokkan teks (text clustering)(Cheng et al., 2022) .

2.2.4 Text Preprocessing

Text Preprocessing ini dilakukan untuk menghasilkan kata yang digunakan sebagai prototype pada setiap dokumen. Setiap kata dicari menutur bentuk dasar kata tersebut menggunakan kamus dasar bahasa Indonesia. Hal ini bertujuan untuk menghindari penyimpanan kata dengan kata dasar sama tetapi dengan imbuhan yang berbeda. Selain menyaring (filtering) kata-kata tidak penting yang memiliki peran sebagai ciri pembeda. Kelompok kata ini biasanya disebut stopword. Pada proses preprocessing terdiri dari beberapa proses yaitu case folding, remove punctuation, stopword, steming, dan tokenizing.

Pada tahap awal preprocessing teks, dilakukan case folding yaitu mengubah karakter huruf besar menjadi huruf kecil. Selain itu, hanya karakter "a" sampai "z" yang akan diterima. Tujuan dari case folding adalah untuk menghindari kesalahpahaman akibat perbedaan huruf besar dan kecil. Proses ini bertujuan untuk mempermudah analisis data oleh program(Zhou et al., 2023) .

Remove punctuation adalah proses yang dilakukan untuk menghapus karakter yang tidak diperlukan seperti tanda baca, angka, markup / html / tag, karakter spesial ($,%, &, dll) atau yang biasa disebut noise. Noise adalah bentuk data yang dapat mengganggu proses pengolahan data. Proses ini bertujuan untuk membersihkan data sebelum dilakukan analisis lebih lanjut(Liu et al., 2020) .

Stopword yaitu tahapan yang berfungsi untuk melakukan Stop atau memberhentikan kata-kata yang dianggap sebagai kata umum, kata ganti, dan kata sambung. Sebagai contoh “Budi berkuliah secara online menggunakan laptop nya setiap hari, sehingga Budi tidak mengenal dekat teman-temannya” akan diubah menjadi “Budi kuliah cara online guna laptop tiap hari hingga Budi tidak kenal dekat teman”(Negara et al., 2019) .

Stemming yaitu tahapan yang bertujuan untuk membuat suatu kata menjadi kata dasar menurut kaidah Bahasa Indonesia yang benar. Sebagai contoh "perkuliahan pada hari ini diundur lagi" akan diubah menjadi "kuliah pada hari ini undur lagi"(Hasan et al., 2021) .

Tokenizing adalah proses membagi atau memotong teks menjadi kata-kata yang membentuknya. Tujuan dari tokenizing adalah untuk mempermudah proses selanjutnya seperti perhitungan kata, pembobotan kata, dan transformasi data menjadi vektor dengan dimensi tinggi. Dengan tokenizing, data teks dapat diolah dengan lebih mudah dan akurat sebelum dilakukan analisis lebih lanjut .

2.2.5 Python

Python adalah bahasa salah satu bahasa pemrograman tingkat tinggi (high-level programming language), berjalan dengan system interpreted, dan bisa dipakai untuk berbagai tujuan (general purpose). Python diciptakan oleh Guido van Rossum pertama kali di Scitchting Mathematisch (CWI) di Belanda pada awal tahun 1990-an .

Python memiliki beberapa fitur dan kelebihan menurut Jubilee Enterprise, yaitu :

1. Memiliki koleksi library yang banyak. Tersedia modul yang siap pakai untuk berbagai kebutuhan.
2. Memiliki struktur bahasa yang jelas dan mudah dipelajari.
3. Python adalah pemrograman yang beorientasi pada objek (OOP). Data dalam Python adalah sebuah objek yang terbuat dari class.
4. Memiliki sistem pengolaan memori otomatis (Gerbage Collection)
5. Bersifat modular, sehingga cukup mudah untuk dikembangkan dengan menciptakan modul baru.

2.2.6 Packages

Python memiliki suatu konsep untuk membungkus beberapa file menjadi kesatuan yang bisa dipanggil (import) ke dalam file lain untuk kebutuhan reusable. Konsep itu dinamakan package dan module. Package adalah kumpulan dari berbagai module, dimana module adalah file Python dengan format .py yang berisikan kumpulan class, fungsi, variabel dan code Python lainnya. Dalam penelitian ini, penulis menggunakan beberapa paket seperti Gensim dan pyLDAvis. Penjelasan mengenai paket tersebut adalah sebagai berikut :

2.2.6.1 Gensin

Gensim adalah library Python bersifat open source untuk pemrosesan bahasa alami. Gensim awalnya dibuat sebagai library pemodelan suatu topik oleh seorang peneliti Ceko bernama Radim Rehurek. Gensim juga dapat melakukan pemrosesan terhadap data teks mentah, Menurut Rehurek (2017) , Gensim memiliki beberapa fitur yang tersedia, yaitu sebagai berikut :

1. Independence memori, yaitu seluruh data corpus tidak perlu dialokasikan di memori RAM karena proses pelatihan terhadap corpus tersebut dilakukan satu per satu.
2. Implementasi efisien dari algoritma ruang vektor, salah satunya adalah Term Frequency-Inverse Document Frequency (TF-IDF).
3. Kemampuan untuk melakukan query kemiripan terhadap dokumen dalam representasi semantiknya.

2.2.6.2 PyLDAvis

PyLDAvis adalah pustaka Python untuk menvisualisasikan model topik interaktif. PyLDAvis dirancang untuk memungkinkan pengguna menafsirkan topik dalam model topik yang sesuai dengan data teks. Paket ini juga mengekstraksi informasi dari model topik LDA untuk menyediakan visualisasi berbasis web interaktif. Visualisasi ini dapat digunakan di dalam notebook IPython, tetapi juga dapat disimpan dalam file HTML untuk memudahkan berbagi dengan pengguna lain .

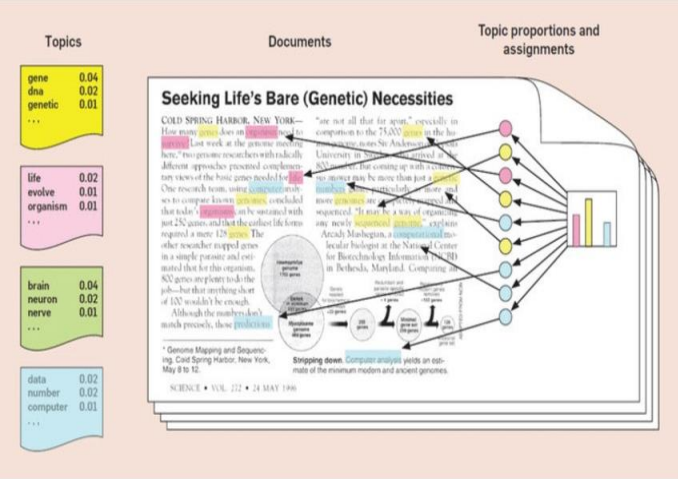
2.2.7 Pemodelan Topik

Pemodelan topik adalah teknik untuk menemukan representasi dokumen sebagai kata kunci dari dokumen. Kata kunci yang ditemukan kemudian digunakan untuk mengindeks dan mencari kembali dokumen sesuai kebutuhan pengguna. Pemodelan topik adalah susunan algoritma untuk menemukan dan memberikan keterangan tematik suatu dokumen untuk menghubungkan beberapa tema dengan entitas yang terkait berdasarkan pembelajaran terpadu menggunakan tema.

2.2.8 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) adalah salah satu metode yang digunakan untuk menganalisis dokumen dengan ukuran yang besar. LDA dapat digunakan untuk meringkas, mengklaster, menghubungkan, atau memproses data yang sangat besar karena LDA menghasilkan daftar topik yang diberi bobot pada setiap dokumen. Distribusi yang digunakan adalah

distribusi Dirichlet, yang digunakan untuk mendapatkan hasil distribusi topik per dokumen. Dalam proses generatif, hasil Dirichlet digunakan untuk mengalokasikan kata-kata dalam dokumen ke topik yang berbeda. LDA beranggapan bahwa dokumen terdiri dari beberapa topik yang memiliki model statistik dari seluruh dokumen. LDA memiliki proses generatif yang menggunakan proses acak terpikirkan pada model yang diasumsikan bahwa dokumen berasal dari topik tertentu. Setiap topik terdiri dari distribusi kata-kata .



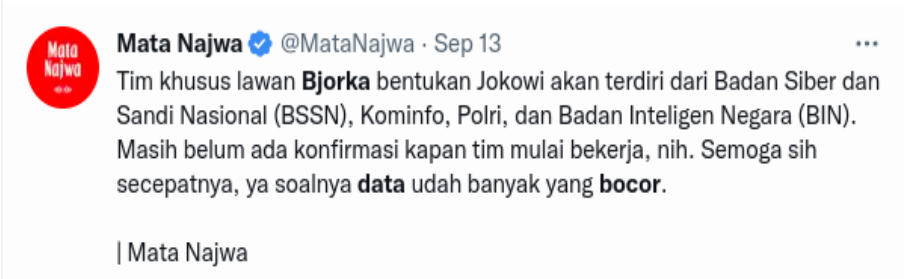
Gambar 2. *Latent Dirichlet Allocation (LDA) (Sumber Blei, D. Probabilistic topic models)*

2.2.9 Topic Coherence

Topic Coherence adalah sebuah metrik yang digunakan untuk menilai seberapa baik satu set kata-kata membentuk topik yang dapat diinterpretasikan dengan mudah oleh manusia. Metrik ini mengukur kesamaan semantik antara kata-kata dalam sebuah topik untuk membantu membedakan topik yang hanya terkait secara statistik dengan topik yang memiliki makna semantik. Topic Coherence adalah suatu ukuran yang digunakan untuk mengevaluasi keberhasilan sebuah model topik. Jika skor koherens topik tinggi, model tersebut dianggap baik. Topik Coherence memberikan interpretasi yang lebih baik daripada Perplexity, meskipun hasil dari matriks perplexity terkadang tidak memiliki korelasi yang baik dengan interpretasi model oleh manusia (Gurcan et al., 2021).

Metode Penelitian

Metode pengumpulan data yang dilakukan dalam penelitian ini yaitu menggunakan pengamatan. Pada tahap ini penulis mengamati dan mengambil data dari Twitter tentang *tweet* atau cuitan masyarakat twitter terhadap keyword data bocor. Penulis mengambil data dengan rentang waktu 01 Septermber 2022 hingga 01 November 2022. Penulis menggunakan *package Twint* dalam pengambilan data yang dilakukan pada Jupyter Notebook. Data yang telah didapatkan kemudian di simpan ke dalam format .CSV. Total data yang penulis ambil sebanyak 11.067 tweet. data tweet tersebut diambil dari banyak akun, diantaranya beberapa akun resmi yaitu Mata Najwa, hariankompas, CNN Indonesia dan sebagainya.



Gambar 3. *Tweet dari akun resmi Mata Najwa*

3.1 Metode *Lifecycle of Text Mining*

Metode yang akan digunakan oleh penulis pada penelitian ini adalah metode *lifecycle of text mining*. Penulis memutuskan untuk menggunakan metode lifecycle of text mining karena untuk menyesuaikan kebutuhan dalam melakukan penelitian ini. Metode ini akan membantu dalam proses dari menemukan data, perencanaan hingga implementasi akhir.

Gambar 4. *Lifecycle of Text Mining Method*

3.2.1 *Problem Definition*

Tahap *problem definition* merupakan proses identifikasi yang harus dilakukan dalam menentukan masalah yang dapat menjadi dasar perumusan dalam penelitian text mining. Pada tahap ini, penulis mendefinisikan masalah sesuai dengan latar belakang yang telah dijelaskan pada pendahuluan (1.1) maka dapat dikembangkan menjadi suatu rumusan masalah yaitu bagaimana mengelompokan topik pembicaraan masyarakat tentang peristiwa kebocoran data di media sosial Twitter menggunakan metode LDA dan bagaimana cara mengambil kesimpulan tentang topik yang ditemukan.

3.2.2 S*election Text Data Mining Approach*

Tahap *selection text mining approach* yaitu suatu proses untuk menentukan metode pendekatan text mining yang tepat sebagai solusi dari masalah yang telah di tentukan pada tahap problem definition. Pada tahap ini penulis menentukan pendekatan text data mining yang digunakan pada penelitian ini. Pendekatan didapatkan dari referensi hasil penelitian sebelumnnya yang telah ditunjukan pada Tabel 2.1. Perbandingan Penelitian Terdahulu. Adapun masalah pada penelitian ini adalah menentukan Topic Modelling dengan Latent Dirichlet Allocation (LDA) tentang peristiwa kebocoran data.

3.2.3 *Data Collection*

Pada tahap *data collection*, penulis mengumpulkan data dari tweet-tweet pada sosial media Twitter dan penulis akan menjadikan data tweet tersebut sebagai input data dalam aplikasi berbasis Python pada penelitian ini. Pengambilan data dilakukan menggunakan cara *Scraping* menggunakan *package Twint* dengan rentang waktu 01 September hingga 01 November. Data yang berhasil didapatkan penulis ada sebanyak 11.067 tweet dengan keyword "data bocor".

3.2.4 *Text Standartization*

Pada tahap *text standardization* dilakukan penyesuaian format data yang bertujuan untuk mempermudah tahap text mining selanjutnya. Data yang telah dikumpulkan pada tahap Data Collection akan dijadikan menjadi satu dengan format yang sesuai standar format data text mining yang telah ditetapkan yaitu menggunakan format .CSV.

3.2.5 *Text Preprocessing*

Pada tahap Text Preprocessing ini, dilakukan pengolahan data untuk mempersiapkan teks menjadi data yang sesuai dengan format yang dibutuhkan dan lebih terstruktur sebelum diproses pada tahap selanjutnya. Pada proses ini dilakukan penyeragaman pada data tweet agar proses text mining selanjutnya lebih mudah dalam pembacaan data tersebut. *Preprocessing* dilakukan dalam beberapa tahapan, yaitu *case folding*, *remove punctuation*, *normalization*, *stopword* dan *tokenizing*.

Tabel 2. *Text Preprocessing* pada *tweet*

Tweet	Case Folding & Remove Punctuation	Stopword & Tokenizing	Normalization
@kemkominfo HALO HALO KOMINFO TANGGUNG JAWAB DONG Itu keamanan nya kekmana, kok bisa sampai bocor data data kita rakyat	halo halo kominfo tanggung jawab dong itu keamanan nya kekmana kok bisa sampai bocor data data kita rakyat indonesia	['halo', 'halo', 'kominfo', 'tanggung', 'aman', 'kekmana', 'rakyat', 'indonesia', 'kasih', 'uang', 'bener']	['halo', 'halo', 'kominfo', 'tanggung', 'aman', 'kekmana', 'rakyat', 'indonesia',

Indonesia bagaimana ini. Bukannya udah dikasih uang banyak ya Masa masih kurang si Yang bener lah	bagaimana ini bukannya udah dikasih uang banyak ya masa masih kurang si yang bener lah		'kasih', 'uang', 'benar']

3.2.6 *Feature Extraction*

Tahap Feature Extraction yaitu tahap pengambilan ciri-ciri yang *unique* dari data yang akan diolah dengan tujuan mengambil informasi yang terpenting dari data dan meningkatkan presisi pengolahan. Penulis menggunakan Bag of Words (BoW) sebagai pembobotan data. Dengan feature extraction akan diperoleh data yang lebih relevan sehingga dapat meningkatkan presisi perhitungan pada tahap selanjutnya.

3.2.7 *Analysis*

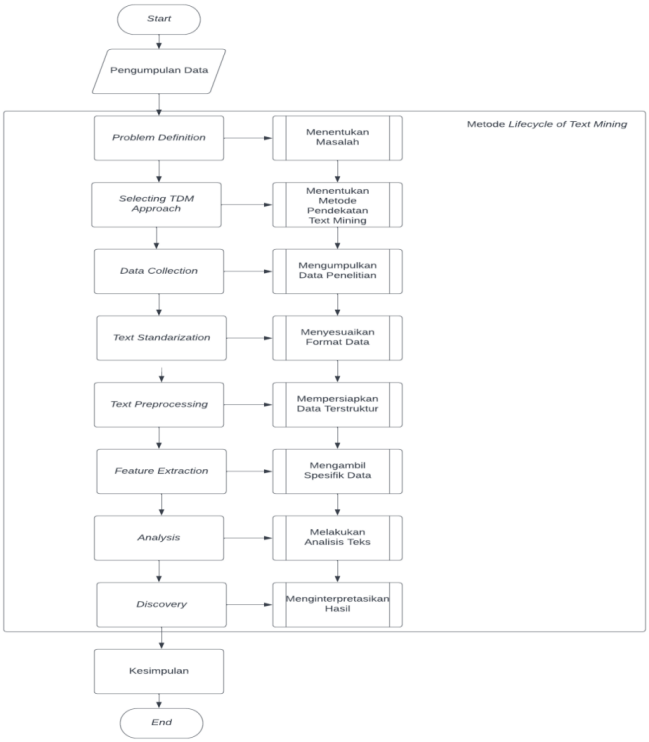
Pada tahap analysis, penulis melakukan Topic Modelling dengan metode Latent Dirichlet Location (LDA) menggunakan aplikasi berbasis Python dengan *package gensim* sebagai pembantu dalam pengamplikasian metode terhadap data yang diolah. Analysis dilakukan setelah melihat nilai *coherence score* topik yang tinggi untuk menentukan berapa jumlah model topik yang akan dipakai sebagai acuan dalam pemodelan. *Coherence Score* dengan nilai tinggi akan mengahasilkan model yang baik.

3.2.8 *Discovery*

Tahap discovery merupakan tahap dimana dilakukan penemuan pola atau pengetahuan dari keseluruhan teks. Dari topik model LDA yang dilakukan akan menghasilkan beberapa topik pembicaraan. Interpreasi selanjutnya akan disajikan ke pengguna dalam bentuk visual.

3.3 Tahapan Penelitian

Dalam penelitian ini dapat digambarkan dalam diagram alir tahapan pemodelan topik sebagai berikut :



Gambar 5. *Tahapan Penelitian*

Hasil dan Pembahasan

4.1 Pengumpulan Data

Penulis mengambil data dengan rentang waktu 01 Septermber 2022 hingga 01 November 2022. Penulis menggunakan package Twint dalam pengambilan data yang dilakukan pada Jupyter Notebook. Data yang telah didapatkan kemudian di simpan ke dalam format .CSV. Total data yang penulis ambil sebanyak 11.067 tweet. Data tweet tersebut diambil dari banyak akun, diantaranya beberapa akun resmi yaitu Mata Najwa, hariankompas, CNN Indonesia dan sebagainya.

Tabel 3. *Dataset Kebocoran Data*

Date	Tweet
9/6/2022 23:57:21	Besok mah yg kerja di kementrian, tes nya harus di ganti tuh. Harus berbasis projek, biar yg kerja pada becus, ga kecolongan mulu. malu atuh, status kementrian kominfo, data bocor mulu.
9/6/2022 22:48:08	KPU Bantah 105 Juta Data Penduduk yang Bocor Berasal dari Mereka https://t.co/0OClaxcvsi https://t.co/M9W9YxkO1A Bedah Data SIM Card yang Bocor: Satu NIK Bisa Seribu Nomor https://t.co/K3NeyKxpgg
9/6/2022 23:57:11	@CNNIndonesia Foto syur cepat direspon, giliran data nik bocor kedodoran ☹

4.2 Text Preprocessing

Dataset yang diambil tidak langsung digunakan pada sistem karena dataset tersebut merupakan data mentah dan unstructured data, maka dilakukan preprocessing untuk meningkatkan kualitas data yang digunakan.

4.2.1 Lowercase Folding dan Remove Punctuation

Data awal merupakan teks yang belum dilakukan pemrosesan apapun, sehingga bentuk dan susunan hurufnya masih tidak teratur. Untuk mengubah semua huruf pada dokumen tersebut menjadi huruf kecil dari a-z maka dilakukan proses Lowercase Folding dan Remove Punctuation untuk menghilangkan simbol simbol yang tidak diperlukan.

Tabel 4. *Lowercase Folding dan Remove Punctuation*

Tweet	Lowercase Folding dan Remove Punctuation
Besok mah yg kerja di Kementrian tes nya harus di ganti tuh harus berbasis projek biar yg kerja pada becus ga kecolongan mulu malu atuh status KEMENTRIAN kominfo data bocor mulu	besok mah yg kerja di kementrian tes nya harus di ganti tuh harus berbasis projek biar yg kerja pada becus ga kecolongan mulu malu atuh status kementrian kominfo data bocor mulu
KPU bantah juta data penduduk yang bocor berasal dari mereka bedah data SIM card yang bocor satu NIK bisa seribunomor	kpu bantah juta data penduduk yang bocor berasal dari mereka bedah data SIM card yang bocor satu NIK bisa seribunomor
Foto syur cepat direspon giliran data NIk bocor kedodoran	foto syur cepat direspon giliran data nik bocor kedodoran

4.2.2 Stopword

Selanjutnya dilakukan proses Stopword untuk memberhentikan kata kata yang dianggap sebagai kata umum, kata ganti dan kata sambung. Karena suatu kalimat yang tidak memiliki makna dianggap kurang bagus dan mempengaruhi keakuratan hasil maka dari itu dilakukan proses Stopword.

Tabel 5. *Stopword*

Tweet	Stopword

Besok mah yg kerja di Kementrian tes nya harus di ganti tuh harus berbasis projek biar yg kerja pada becus ga kecolongan mulu malu atuh status KEMENTRIAN kominfo data bocor mulu	Besok kerja di Kementrian tes harus di ganti harus berbasis projek biar kerja pada kecolongan malu status KEMENTRIAN kominfo data bocor
KPU bantah juta data penduduk yang bocor berasal dari mereka bedah data SIM card yang bocor satu NIK bisa seribunomor	KPU bantah juta data pendudukbocor berasal mereka bedah data SIM card bocor satu NIK seribunomor
Foto syur cepat direspon giliran data NIk bocor kedodoran	Foto syur cepat direspon giliran data NIk bocor kedodoran

4.2.3 Tokenizing

Data dari hasil lowercase folding dan juga remove punctuation selanjutnya akan dilakukan proses tokenizing yaitu melakukan pembagian atau pemotongan teks menjadi kata per kata untuk mempermudah proses perhitungan kata, pembobotan kata, dan transformasi data mrnjadi vektor.

Tabel 6. *Tokenizing*

Tweet	Tokenizing
besok mah yg kerja di kementrian tes nya harus di ganti tuh harus berbasis projek biar yg kerja pada becus ga kecolongan mulu malu atuh status kementrian kominfo data bocor mulu	['besok', 'mah', 'yg', 'kerja', 'di', 'kementrian', 'tes', 'nya', 'harus', 'di', 'ganti', 'tuh', 'harus', 'berbasis', 'projek', 'biar', 'yg', 'kerja', 'pada', 'becus', 'ga', 'kecolongan', 'mulu', 'malu', 'atuh', 'status', 'kementrian', 'kominfo', 'data', 'bocor', 'mulu']
kpu bantah juta data penduduk yang bocor berasal dari mereka bedah data SIM card yang bocor satu NIK bisa seribunomor	['kpu', 'bantah', 'juta', 'data', 'penduduk', 'yang', 'bocor', 'berasal', 'dari', 'mereka', 'bedah', 'data', 'sim', 'card', 'yang', 'bocor', 'satu', 'nik', 'bisa', 'seribunomor']
foto syur cepat direspon giliran data nikbocor kedodoran	['foto', 'syur', 'cepat', 'direspon', 'giliran', 'data', 'nik', 'bocor', 'kedodoran']

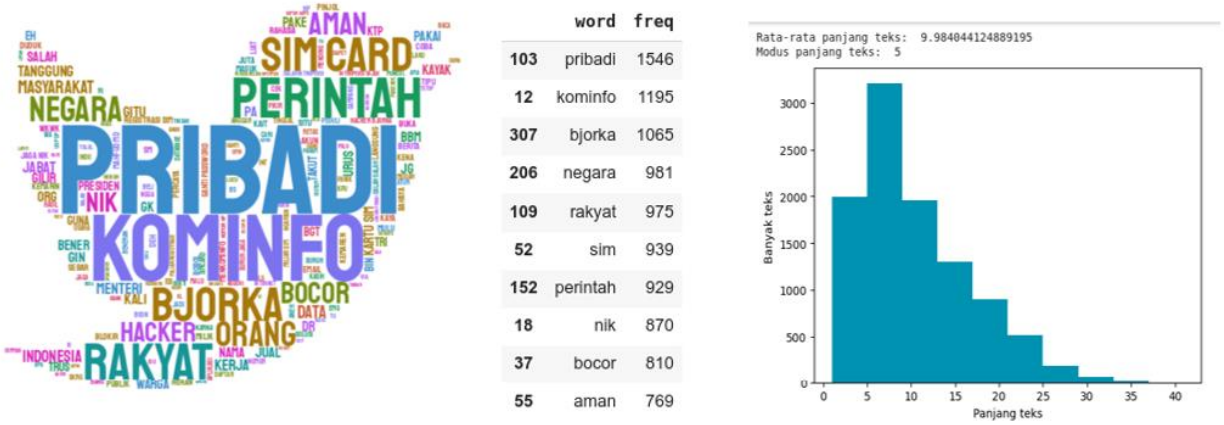
3.2.4 Stemming

Setelah itu, data diolah dengan tahap stemming untuk membuat suatu kata menjadi kata dasar menurut kaidah Bahasa Indonesia yang benar. Stemming dilakukan untuk menghilangkan kata imbuhan yang terdapat pada kalimat dan juga menghilangkan prefiks serta sufiks berdasarkan Bahasa Indonesia yang baik dan benar

Tabel 7. *Stemming*

Tweet	Stemming
besok mah yg kerja di kementrian tes nya harus di ganti tuh harus berbasis projek biar yg kerja pada becus ga kecolongan mulu malu atuh status kementrian kominfo data bocor mulu	kerja tri tes ganti bas projek kerja becus colong mulu malu atuh status tri kominfo mulu
kpu bantah juta data penduduk yang bocor berasal dari mereka bedah data SIM card yang bocor satu NIK bisa seribunomor	kpu bantah juta duduk asal bedah sim card nik seribunomor
foto syur cepat direspon giliran data nikbocor kedodoran	foto syur cepat respon gilir nik kedodoran

Data yang sudah diolah dapat divisualisasikan melalui wordcloud. Visualisasi dapat memperlihatkan kata kata yang dicetak dalam ukuran besar dan kecil yang merepresentasikan frekuensi munculnya kata.



	word	freq
103	pribadi	1546
12	kominfo	1195
307	bjorka	1065
206	negara	981
109	rakyat	975
52	sim	939
152	perintah	929
18	nik	870
37	bocor	810
55	aman	769

Gambar 6. *Hasil Preprocessing*

Dari hasil preprocessing didapat kata 'pribadi' menjadi kata terbanyak dengan frekuensi 1546 kemunculan kata, lalu 'kominfo' dengan frekuensi 1195 kemunculan kata dan 'bjorka' dengan frekuensi 1065 kemunculan kata. Rata-rata panjang teks yang didapat yaitu 9.98 kata dengan panjang teks 5.

3.3 Feature Extraction

Tahap ini merupakan tahap pengambilan ciri ciri yang unik dari data hasil pemrosesan sebelumnya dan juga untuk meningkatkan tingkat akurasi terhadap data tersebut. Metode yang dilakukan pada tahap ini yaitu menggunakan metode *Latent Dirichlet Allocation* untuk pemodelan topiknya. Pada penelitian ini, score coherences dihitung dengan memberi batasan sebanyak 5, dimulai dengan 1 topik terbaik yang diintrepretasikan dan diiterasi sebanyak 40 kali.

3.3.1 Bag of Word (BoW)

Model Bag of Word dilakukan untuk mengesktraksi teks dan beberapa fitur dari teks, Metode ini digunakan untuk menghitung kemunculan setiap kata pada suatu dokumen sehingga memudahkan dalam melihat frekuensi kemunculan kata

```
from pprint import pprint
pprint(coherences)
```

```
[0.4589483707911217,
 0.43796724994722813,
 0.4728355056964541,
 0.41328821898115087,
 0.4100343095293712]
```

Gambar 7. *Score Coherences dengan BoW*

Percobaan dilakukan dengan jumlah topik dari 1 sampai jumlah topik 5. Dengan fitur ekstraksi BoW menghasilkan score coherences yang baik rata-rata diangka 41-47, fitur ini menunjukan hasil yang terbaik pada jumlah topik 3, berikut ini adalah topik dengan score coherences terbaik.

```
[(0,
  '0.013*"orang" + 0.012*"rakyat" + 0.011*"perintah" + 0.010*"nik" + '
  '0.010*"pribadi" + 0.007*"kominfo" + 0.007*"jaga" + 0.007*"masyarakat" + '
  '0.007*"data" + 0.006*"kerja"'),
 (1,
  '0.035*"johnny" + 0.034*"bjorka" + 0.027*"plate" + 0.024*"pribadi" + '
  '0.023*"bocor" + 0.022*"hacker" + 0.021*"menkominfo" + 0.018*"indonesia" + '
  '0.015*"negara" + 0.014*"sim"'),
 (2,
  '0.025*"lindung" + 0.017*"negara" + 0.017*"pdp" + 0.014*"pribadi" + '
  '0.014*"uu" + 0.010*"bbm" + 0.009*"rakyat" + 0.009*"ruu" + 0.009*"sistem" + '
  '0.007*"digital"')]
```

Gambar 8. *Isi Topik dengan Bow*

Gambar 9. *Visualisasi Pyldavis BoW*

Topik pertama berisi ‘orang’, ‘rakyat’, ‘perintah’, ‘nik’, ‘pribadi’, ‘kominfo’, ‘jaga’, ‘masyarakat’, ‘data’, ‘kerja’. Menginterpretasikan topik tentang kominfo harus menjaga data pribadi seperti NIK rakyat. Topik kedua berisi ‘johnny’, ‘bjorka’, ‘plate’, ‘pribadi’, ‘bocor’, ‘hacker’, ‘menkominfo’, ‘indonesia’, ‘negara’, ‘sim’ menginterpretasikan topik tentang johnny g plate selaku menkominfo di indonesia bertanggung jawab atas kasus kebocoran data ulah hacker bjorka. Topik ketiga berisi ‘lindung’, ‘negara’, ‘pdp’, ‘pribadi’, ‘uu’, ‘bbm’, ‘rakyat’, ‘ruu’, ‘sistem’, ‘digital’ menginterpretasikan topik tentang perlindungan data pribadi rakyat memalui ruu pdp.

3.3.2 TF-IDF

Pembobotan term atau kata pada dokumen dimulai dari perhitungan term frequency. Kemunculan setiap kata di dalam satu dokumen di dataset ini dihitung dan dinyatakan dalam nilai desimal. Setelah itu, pembobotan dari IDF menghasilkan nilai yang sama seperti TF yaitu nilai desimal. TF-IDF sering digunakan sebagai skema pembobotan untuk menentukan pentingnya sebuah kata dalam sebuah dokumen atau korpus. Semakin tinggi nilai TF-IDF, semakin penting kata tersebut dianggap.

```
from pprint import pprint
pprint(coherences)
```

```
[0.2997090818502128,
 0.34120565630733674,
 0.22779837548096019,
 0.4162606079798459,
 0.4797712177878327]
```

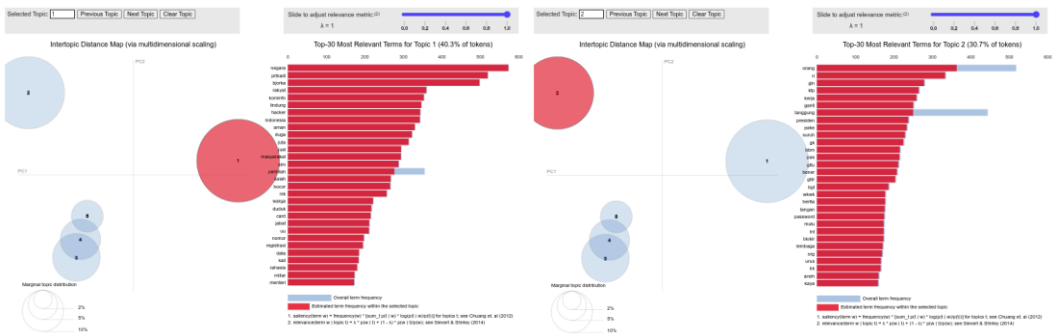
Gambar 10. *Score Coherences dengan TF-IDF*

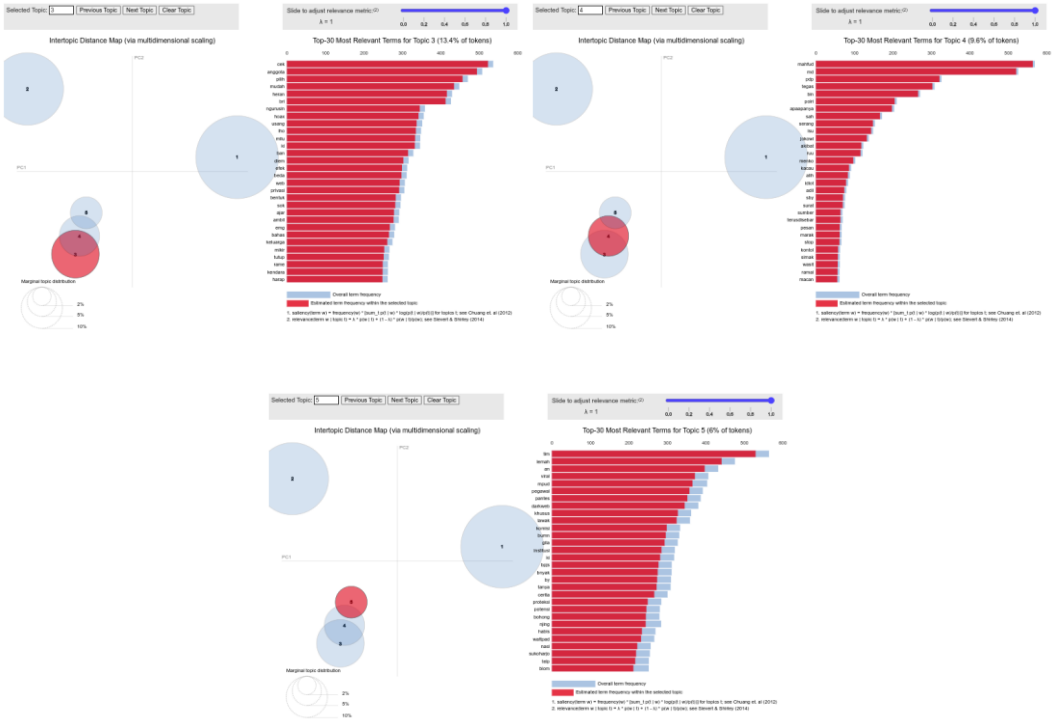
Percobaan dilakukan pada fitur ekstraksi ini sama dengan sebelumnya, dengan jumlah topik dari 1 sampai jumlah topik 5. Dengan fitur ekstraksi TF-IDF menghasilkan score coherences yang tidak rata diangka 22-47, fitur ini menunjukan hasil yang terbaik pada jumlah topik 5, berikut ini adalah topik dengan score coherences terbaik.

```
[(0,
  '0.014*"negara" + 0.013*"pribadi" + 0.012*"bjorka" + 0.009*"rakyat" + '
  '0.009*"kominfo" + 0.008*"lindung" + 0.008*"hacker" + 0.008*"indonesia" + '
  '0.008*"aman" + 0.008*"duga"'),
 (1,
  '0.021*"mahfud" + 0.020*"md" + 0.012*"pdp" + 0.011*"tegas" + 0.010*"bin" + '
  '0.008*"polri" + 0.007*"apaapanya" + 0.006*"sah" + 0.006*"serang" + '
  '0.005*"isu"'),
 (2,
  '0.007*"orang" + 0.006*"ri" + 0.005*"gin" + 0.005*"ktp" + 0.005*"kerja" + '
  '0.005*"ganti" + 0.005*"tanggung" + 0.004*"presiden" + 0.004*"pake" + '
  '0.004*"suruh"'),
 (3,
  '0.005*"cek" + 0.005*"anggota" + 0.004*"pilih" + 0.004*"mudah" + '
  '0.004*"heran" + 0.004*"bri" + 0.003*"ngurusin" + 0.003*"hoax" + '
  '0.003*"usang" + 0.003*"lho"'),
 (4,
  '0.005*"tim" + 0.004*"lemah" + 0.004*"an" + 0.004*"viral" + 0.004*"mpud" + '
  '0.003*"pegawai" + 0.003*"pantes" + 0.003*"darkweb" + 0.003*"khusus" + '
  '0.003*"lawak"')]
```

Gambar 11. *Isi Topik dengan TF-IDF*

Topik pertama berisi ‘negara’, ‘pribadi’, ‘bjorka’, ‘rakyat’, ‘kominfo’, ‘lindung’, ‘hacker’, ‘indonesia’, ‘aman’, ‘duga’. Menginterpretasikan topik tentang kominfo lindungi data pribadi dari hacker bjorka. Topik kedua berisi ‘mahfud’, ‘md’, ‘pdp’, ‘tegas’, ‘bin’, ‘polri’. ‘apaapanya’, ‘sah’, ‘serang’, ‘isu’, isi dari topik kedua ini susah untuk diintrepretasikan karena kata kata yang dihasilkan tidak saling berhubungan. Untuk topik ketiga sampai topik kelima juga susah untuk diinterpretasikan.





Gambar 12. *Visualisasi Pyldavis TF-IDF*

3.4 Analisis Perbandingan BoW dan TF IDF

Dari data-data diatas didapatkan perbandingan antara menggunakan fitur ekstraksi BoW dan fitur ekstraksi TF-IDF sebagai berikut. Berdasarkan Score Coherences :

Tabel 8. *Score Coherences TF-IDF dan BOW*

Jumlah Topik	BoW	TF-IDF
1	0.4589483707911217	0.2997090818502128
2	0.43796724994722813	0.34120565630733674
3	0.4728355056964541	0.22779837548096019
4	0.41328821898115087	0.4162606079798459
5	0.4100343095293712	0.4797712177878327

Data BoW mendapatkan score coherences dengan rata rata nilai 40 keatas sehingga kata-kata yang dihasilkan dengan fitur ekstraksi BoW lebih mudah diinterpretasikan oleh bahasa manusia. Sedangkan TF-IDF mendapat score coherences mulai dari 0.22, 0.29, 0.34, 0.41, dan 0.47. Dari hasil ini BoW menjadi fitur ekstraksi terbaik untuk pemodelan topik menggunakan metode *Latent Diriclet Allocation*.

Berdasarkan isi topik, perbandingan dilakukan dengan memilih score coherences terbaik pada masing-masing fitur ekstraksi, fitur ekstraksi BoW terbaik di jumlah topik 3 dengan

score coherences 0.47 dan fitur ekstraksi TF-IDF terbaik di jumlah topik 5 dengan score cohereces 0.47 sebagai berikut :

Tabel 9. *Hasil Topik BOW dan TF-IDF*

Jumlah Topik	BoW	TF-IDF
1	'orang', 'rakyat', 'perintah', 'nik', 'pribadi', 'kominfo', 'jaga', 'masyarakat', 'data', 'kerja'	'negara', 'pribadi', 'bjorka', 'rakyat', 'kominfo', 'lindung', 'hacker', 'indonesia', 'aman', 'duga'
2	'johnny', 'bjorka', 'plate', 'pribadi', 'bocor', 'hacker', 'menkominfo', 'indonesia', 'negara', 'sim'	'mahfud', 'md', 'pdp', 'tegas', 'bin', 'polri'. 'apaapanya', 'sah', 'serang', 'isu'
3	'lindung', 'negara', 'pdp', 'pribadi', 'uu', 'bbm', 'rakyat', 'ruu', 'sistem', 'digital'	'orang', 'ri', 'gin', 'ktp', 'kerja', 'ganti', 'tanggung', 'presiden', 'pake', 'suruh'
4		'cek', 'anggota', 'pilih', 'mudah', 'heran', 'bri', 'ngurusin', 'hoax', 'usang', 'lho'
5		'tim', 'lemah', 'an', 'viral', 'mpud', 'pegawai', 'pantes', 'darkweb', 'khusus', 'lawak'

Dari perbandingan isi topik pada tabel diatas didapatkan bahwa isi topik menggunakan fitur ekstraksi BoW lebih baik daripada menggunakan fitur ekstraksi TF-IDF karena kata kata yang dihasilkan saling terhubung sehingga lebih mudah di interpretasikan oleh bahasa manusia dengan baik.

Simpulan

Pada penelitian ini telah diaplikasikan fitur ekstraksi BoW dan fitur ekstraksi TF-IDF pada pemodelan topik LDA terhadap 11.067 data tentang kebocoran data di Twitter yang diambil penulis dengan rentang waktu waktu 01 September 2022 hingga 01 November 2022. Dengan perbandingan fitur ekstraksi BoW dan TF-ID didapatkan hasil fitur ekstraksi BoW

lebih baik daripada fitur ekstraksi TF-IDF. Fitur ekstraksi BoW mendapatkan score coherences yang tinggi dan isi topik yang dihasilkan juga bisa diinterpretasikan dengan mudah seperti topik 2 berisi menkominfo johnny g plate menjadi sorotan warga twitter terkait bocornya data SIM ulah *hacker Bjorka*.

Daftar Pustaka

Aditya, B. R. (2015). Penggunaan web crawler untuk menghimpun tweets dengan metode pre-processing text mining. Jurnal INFOTEL - Informasi Telekomunikasi Elektronika, 7(2), 93. https://doi.org/10.20895/infotel.v7i2.35

Agustina, D. A., Subanti, S., & Zukhronah, E. (2021). Implementasi text mining pada analisis sentimen pengguna Twitter terhadap marketplace di Indonesia menggunakan algoritma Support Vector Machine. Indonesian Journal of Applied Statistics, 3(2), 109. https://doi.org/10.13057/ijas.v3i2.44337

Akhir, T. (2018). Similarity berbasis web responsive.

Arifianto, E. Y., & Program, K. F. D. (2020). Analisis topik data tindak kriminal pada media sosial Twitter menggunakan metode LDA (Latent Dirichlet Allocation).

Cheng, X., Cao, Q., & Liao, S. S. (2022). An overview of literature on COVID-19, MERS and SARS: Using text mining and latent Dirichlet allocation. Journal of Information Science. https://doi.org/10.1177/0165551520954674

Daftar Pustaka

Fauziyyah, A. K. (2020). Analisis sentimen pandemi Covid19 pada streaming Twitter dengan text mining Python. Jurnal Ilmiah SINUS, 18(2), 31. https://doi.org/10.30646/sinus.v18i2.491

Gharehchopogh, F. S., & Khalifehlou, Z. A. (2011). Analysis and evaluation of unstructured data: Text mining versus natural language processing. In 2011 International Conference on Advanced ICT (ICAICT) (pp. 1-5). IEEE. https://doi.org/10.1109/ICAICT.2011.6111017

Gurcan, F., Ozyurt, O., & Cagitay, N. E. (2021). Investigation of emerging trends in the e-learning field using latent dirichlet allocation. … Review of Research in Open and …. https://www.erudit.org/en/journals/irrodl/2021-v22-n2-irrodl06128/1078397ar/abstract/

Hadi, A. F., W, D. B. C., Hasan, M., & Penelitian, A. D. (2017). Text mining pada media sosial Twitter studi kasus: Masa tenang Pilkada DKI.

Han, X. (2020). Evolution of research topics in LIS between 1996 and 2019: An analysis based on latent Dirichlet allocation topic model. Scientometrics. https://doi.org/10.1007/s11192-020-03721-0

Hasan, M., Rahman, A., Karim, M. R., Khan, M. S. I., & ... (2021). Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA). … Conference on Trends …. https://doi.org/10.1007/978-981-33-4673-4_27

Islam, U., Syarif, N., Studi, P., Informatika, T., Sains, F., & Teknologi, D. A. N. (2021). Latent Dirichlet Allocation (LDA) untuk mengetahui topik pembicaraan warganet.

Janmaijaya, M., Shukla, A. K., Muhuri, P. K., & ... (2021). Industry 4.0: Latent Dirichlet Allocation and clustering based theme identification of bibliography. … Applications of Artificial …. https://www.sciencedirect.com/science/article/pii/S0952197621001275

John, J. S., John, K. S., & Han, B. (2022). Entrepreneurial crowdfunding backer motivations: a latent Dirichlet allocation approach. European Journal of Innovation …. https://doi.org/10.1108/EJIM-05-2021-0248

Kabiru, I. N., & Sari, P. K. (2019). Analisa konten media sosial e-commerce pada Instagram menggunakan metode sentiment analysis dan LDA-based topic modeling (studi kasus: Shopee Indonesia). eProceedings Manajemen, 6(1), 12-19. https://openlibrarypublications.telkomuniversity.ac.id/index.php/management/article/view/8498

Kang, H. J., Kim, C., & Kang, K. (2019). Analysis of the trends in biochemical research using latent Dirichlet allocation (LDA). Processes. https://www.mdpi.com/2227-9717/7/6/379

Komputer, I., Ilmu, D., Matematik, F., Alam, P., & Mada, U. G. (2018). 5936-Article Text-8497-11551-10-20181123, 5(September).

Kozlowski, D., Semeshenko, V., & Molinari, A. (2021). Latent Dirichlet allocation model for world trade analysis. PloS One. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0245393

Kusnadi, S. A. (2021). Perlindungan hukum data pribadi sebagai hak privasi. Al Wasath: Jurnal Ilmu Hukum, 2(1), 9-16. https://doi.org/10.47776/alwasath.v2i1.127

Liu, Y., Du, F., Sun, J., & Jiang, Y. (2020). iLDA: An interactive latent Dirichlet allocation model to improve topic quality. Journal of Information Science. https://doi.org/10.1177/0165551518822455

Negara, E. S., Triadi, D., & Andryani, R. (2019). Topic modelling twitter data with latent dirichlet allocation method. … International Conference on …. https://ieeexplore.ieee.org/abstract/document/8984523/

Putra, I. M. K. B., & Kusumawardani, R. P. (2017). Analisis topik informasi publik media sosial di Surabaya menggunakan pemodelan Latent Dirichlet Allocation (LDA) [Topic analysis of public information in social media in Surabaya based on Latent Dirichlet Allocation (LDA) topic modeling]. Jurnal Teknologi Informasi dan Sistem, 6(2), 2-7.

Rahman, M. N. (2022). Analisis performa penggunaan stopwords dan stemming dalam sentimen analisis dengan pendekatan klasifikasi naive bayes.

Tran, B. X., Latkin, C. A., Sharafeldin, N., & ... (2019). Characterizing artificial intelligence applications in cancer research: a latent dirichlet allocation analysis. JMIR Medical …. https://medinform.jmir.org/2019/4/e14401/

Wang, Y., Tong, Y., & Shi, D. (2020). Federated latent dirichlet allocation: A local differential privacy based framework. Proceedings of the AAAI Conference on …. https://ojs.aaai.org/index.php/AAAI/article/view/6096

Watrianthos, R., Giatman, M., Simatupang, W., Syafriyeti, R., & Daulay, N. K. (2022). Analisis sentimen pembelajaran campuran menggunakan data Twitter. Jurnal Media Informasi Budidarma, 6(1), 166. https://doi.org/10.30865/mib.v6i1.3383

Wirasakti, L. A., Permadi, R., Hartanto, A. D., & Hartatik, H. (2020). Pembuatan kata kunci otomatis dalam artikel dengan pemodelan topik. Jurnal Media Informasi Budidarma, 4(1), 27. https://doi.org/10.30865/mib.v4i1.1707

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., & Zhu, T. (2020). Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. PloS One. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0239441

Zhou, S., Kan, P., Huang, Q., & ... (2023). A guided latent Dirichlet allocation approach to investigate real-time latent topics of Twitter data during Hurricane Laura. Journal of Information …. https://doi.org/10.1177/01655515211007724