SKRIPSI

PENERAPAN *TOPIC MODELING* MENGGUNAKAN *LATENT DIRICHLET ALLOCATION* (LDA) DALAM BERITA KESEHATAN GIZI

Disusun dan diajukan oleh:

HADRIANA NURUL PERTIWI
D121 20 1059

PROGRAM STUDI SARJANA TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS HASANUDDIN
GOWA
2024

LEMBAR PENGESAHAN SKRIPSI

PENERAPAN TOPIC MODELING MENGGUNAKAN LATENT DIRICHLET ALLOCATION (LDA) DALAM BERITA KESEHATAN GIZI

Disusun dan diajukan oleh

Hadriana Nurul Pertiwi
D121 20 1059

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Program Studi Teknik Informatika
Fakultas Teknik Universitas Hasanuddin
Pada tanggal 21 Agustus 2024
dan dinyatakan telah memenuhi syarat kelulusan

Menyetujui,

Pembimbing Utama,

Elly Warni, S.T., M.T.
NIP 19820216 200812 2 0001

Ketua Program Studi,

Prof. Dr. Ir. Indrabayu, S.T., M.T., M.Bus.Sys., IPM., ASEAN.Eng.
NIP 19750716 200212 1 004

PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini ;
Nama : Hadriana Nurul Pertiwi
NIM : D121 20 1059
Program Studi : Teknik Informatika
Jenjang : S1

Menyatakan dengan ini bahwa karya tulisan saya berjudul

Penerapan *Topic Modeling* Menggunakan *Latent Dirichlet Allocation* (LDA) Dalam Berita Kesehatan Gizi

Adalah karya tulisan saya sendiri dan bukan merupakan pengambilan alihan tulisan orang lain dan bahwa skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri.

Semua informasi yang ditulis dalam skripsi yang berasal dari penulis lain telah diberi penghargaan, yakni dengan mengutip sumber dan tahun penerbitannya. Oleh karena itu semua tulisan dalam skripsi ini sepenuhnya menjadi tanggung jawab penulis. Apabila ada pihak manapun yang merasa ada kesamaan judul dan atau hasil temuan dalam skripsi ini, maka penulis siap untuk diklarifikasi dan mempertanggungjawabkan segala resiko.

Segala data dan informasi yang diperoleh selama proses pembuatan skripsi, yang akan dipublikasi oleh Penulis di masa depan harus mendapat persetujuan dari Dosen Pembimbing.

Apabila dikemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan isi skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Makassar, 07 Oktober 2024

Yang Menyatakan

Hadriana Nurul Pertiwi

ABSTRAK

HADRIANA NURUL PERTIWI. *Penerapan Topic Modeling Menggunakan Latent Dirichlet Allocation (LDA) Dalam Berita Kesehatan Gizi* (dibimbing oleh Elly Warni)

Dalam era informasi digital yang berkembang pesat, berita kesehatan gizi menjadi sumber penting untuk informasi terbaru mengenai isu-isu gizi. Jumlah berita kesehatan gizi *online* yang terus meningkat menimbulkan tantangan baru dalam menganalisis topik utama karena proses manual memerlukan banyak waktu. Selain itu, melacak perkembangan atau tren topik gizi dari waktu ke waktu juga sulit. Oleh karena itu, diperlukan sistem untuk menemukan topik utama dan tren dari berita kesehatan gizi.

Penelitian ini bertujuan menerapkan *topic modeling* dalam berita kesehatan gizi menggunakan *Latent Dirichlet Allocation* (LDA) dan menganalisis kinerjanya dalam mengelompokkan tren topik.

Data diambil dari berita kesehatan gizi di portal berita *online* Indonesia, di-scraping berdasarkan tahun 2019 hingga 2023. Tahapannya meliputi *preprocessing*, *feature extraction*, *topic modeling*, dan analisis hasil. Implementasi *preprocessing* mencakup *case folding*, *punctuation removal*, *number removal*, *stopword removal*, *stemming*, dan *tokenizing*. *Feature extraction* melibatkan pembobotan kata dengan *Term Frequency-Inverse Document Frequency* (TF-IDF) serta *bigram* dan *trigram*. *Topic modeling* dilakukan dengan menggunakan *corpus* dan *dictionary*, metode LDA, dan evaluasi model dengan topic coherence.

Hasil penelitian menunjukkan bahwa LDA dapat mengidentifikasi sejumlah topik utama dalam berita dengan jumlah topik yang terbentuk pada tahun 2019, 2020, 2021, 2022 dan 2023 secara berturut-turut adalah 9, 10, 4, 8 dan 8 topik. Tahun 2019 dengan nilai *coherence* 0.42057 mencakup nutrisi anak dan manajemen kesehatan serta pola makan dan risiko kesehatan. Tahun 2020 dengan nilai *coherence* 0.4656 berfokus pada nutrisi ibu hamil dan anak. Tahun 2021 dengan nilai *coherence* 0.45326 mengenai kesehatan dan gizi selama pandemi. Tahun 2022 dengan nilai *coherence* 0.41281 mencakup manajemen kesehatan dan hidrasi selama puasa. Tahun 2023 dengan nilai *coherence* 0.46266 berfokus pada gizi dan keamanan pangan.

Kata Kunci: Berita kesehatan gizi, *topic modeling*, *Latent Dirichlet Allocation*, analisis tren

ABSTRACT

HADRIANA NURUL PERTIWI. *Application of Topic Modeling Using Latent Dirichlet Allocation (LDA) in Nutrition Health News* (supervised by Elly Warni)

In the era of rapidly growing digital information, nutritional health news has become an important source of up-to-date information on nutrition issues. The ever-increasing number of online nutritional health news poses new challenges in analyzing key topics as manual processes require a lot of time. Moreover, tracking the development or trends of nutrition topics over time is also difficult. Therefore, a system is needed to discover the main topics and trends of nutritional health news.

This study aims to apply topic modeling in nutritional health news using Latent Dirichlet Allocation (LDA) and analyze its performance in clustering topic trends.

Data was taken from nutritional health news on Indonesian online news portals, scraped based on 2019 to 2023. The stages include preprocessing, feature extraction, topic modeling, and result analysis. The preprocessing implementation includes case folding, punctuation removal, number removal, stopword removal, stemming, and tokenizing. Feature extraction involves word weighting with Term Frequency-Inverse Document Frequency (TF-IDF) as well as bigrams and trigrams. Topic modeling is done using corpus and dictionary, LDA method, and model evaluation with topic coherence.

The results show that LDA can identify a number of main topics in the news with the number of topics formed in 2019, 2020, 2021, 2022 and 2023 are 9, 10, 4, 8 and 8 topics respectively. 2019 with a coherence value of 0.42057 includes child nutrition and health management and diet and health risks. 2020 with a coherence value of 0.4656 focuses on maternal and child nutrition. 2021 with a coherence value of 0.45326 on health and nutrition during the pandemic. Year 2022 with a coherence value of 0.41281 covers health management and hydration during fasting. Year 2023 with a coherence value of 0.46266 focuses on nutrition and food safety.

Keywords: Nutrition health news, topic modeling, Latent Dirichlet Allocation, trend analysis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR SINGKATAN DAN ARTI SIMBOL

Lambang/Singkatan	Arti dan Keterangan
α	Hyperparameter *Dirichlet* untuk distribusi topik dalam dokumen
θ	Probabilitas *i* dokumen yang mengandung *j* topik
β	hyperparameter *Dirichlet* untuk distribusi kata dalam topik
M	Jumlah dokoumen
N	Jumlah kata dalam setiap dokumen
w	Kata dalam dokumen
z	Topik yang terbentuk
φ	Distribusi kata terhadap topik dalam corpus
d	Dokumen
TF	*Term Frequency*
DF	*Document Frequency*
IDF	*Inverse Document Frequency*
BoW	*Bag-of-Words*
LDA	*Latent Dirichlet Allocation*
ND	jumlah setiap topik dalam dokumen
NW	jumlah setiap kata dalam topik
CV	*Coherence Value*
PMI	*Pointwise Mutual Information*
MDS	*Multidimensional Scaling*
JSD	*Jensen-Shannon Divergence*
KL	*Kullback-Leibler*
PC	*Principal Component*
SDGs	*Sustainable Development Goals*

DAFTAR LAMPIRAN

KATA PENGANTAR

Puji syukur kehadirat Allah SWT atas berkat rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan tugas akhir ini yang berjudul "Penerapan Topic Modeling Menggunakan Latent Dirichlet Allocation (LDA) Dalam Berita Kesehatan Gizi" sebagai salah satu persyaratan yang harus dipenuhi dalam menyelesaikan jenjang Strata-1 pada Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin. Sholawat serta salam kepada nabi Muhammad SAW yang telah menunjukkan dan mengajarkan akhlak mulia sehingga didapatkan kenyamanan dan keramahan dalam berhubungan dengan orang di sekitar.

Dengan segala kerendahan hati, penulis menyadari bahwa dalam penyusunan dan penulisan laporan tugas akhir ini tidak lepas dari bantuan, bimbingan serta dukungan dari berbagai pihak, dari masa perkuliahan sampai dengan masa penyusunan tugas akhir ini. Sehingga, pada kesempatan ini penulis mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Allah SWT. Atas semua karunia serta pertolongan-Nya yang tiada batas, yang telah diberikan kepada penulis di setiap langkah salam penelitian hingga penulisan laporan ini.
2. Kedua orang tua penulis, Bapak Husain Gafur dan Ibu Lenny yang selalu mendoakan untuk kebaikan penulis, selalu memberikan kasih sayang, cinta, dukungan dan motivasi. Terima kasih Bapak dan Ibu yang selalu sabar dan semangat tiada hentinya dalam menghadapi dan mendidik penulis.
3. Ibu Elly Warni, S.T., M.T., selaku pembimbing utama yang senantiasa meluangkan waktu, tenaga dan pikiran serta perhatian yang luar biasa untuk membimbing penulis dalam proses penyelesaian tugas akhir ini.
4. Segenap Dosen dan Staf Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin yang telah memberikan ilmu dan pengetahuan baru, serta bantuan kepada penulis selama menuntut masa perkuliahan.
5. Teman-teman REZOLVER yang telah banyak membantu selama kuliah.

6. Miqa, Paula, Mutiah, May, Thesa, Fadhlu, Reiky yang telah membersamai penulis, yang selalu menemani penulis dan memberi masukan selama menyelesaikan tugas akhir. Terima kasih sudah ingin menjadi teman penulis di masa perkuliahan ini.
7. Ainun, Dzaky, Adil, Yahdi, Ali, Adit, Iman, Ravi yang menjadi teman pertama penulis saat masuk perkuliahan sampai sekarang.
8. Ghina, Nure, Kiki yang telah menemani, menghibur, dan ingin direpotkan sama penulis selama mengerjakan tugas akhir ini.
9. Sahabat penulis Dhea Putri Utami yang selalu menyemangati dan menjadi tempat berkeluh kesah penulis selama ini.
10. Tak lupa kepada penulis sendiri, terima kasih telah bertahan sejauh ini, menyelesaikan tugas akhir walaupun banyak rintangan dan tantangan yang dihadapi. Semoga setelah menyelesaikan semuanya bisa mendapatkan hal-hal yang baik di luar sana dan dilancarkan segala urusannya.
11. Serta pihak-pihak lain yang tidak sempat disebutkan dan tanpa sadar telah membantu penulis dalam menyelesaikan tugas akhir.

Penulis memberikan rasa hormat yang tak terhingga, semoga Allah SWT. membalas semua kebaikan dari semua pihak yang telah banyak membantu penulis dalam menyelesaikan tugas akhir ini. Penulis menyadari tugas akhir ini masih banyak kekurangan dan jauh dari kata sempurna. Oleh karena itu, penulis mengharapkan segala masukan dan saran yang membangun sehingga tugas akhir ini dapat memberi manfaat bagi penulis dan pembaca. Akhir kata, semoga tugas akhir ini dapat dijadikan sebagai sumber ilmu pengetahuan dan bermanfaat bagi penulis dan pembaca pada umumnya.

Makassar, 07 Oktober 2024

Hadriana Nurul Pertiwi

BAB I
PENDAHULUAN

1.1 Latar Belakang

Dalam era informasi digital yang berkembang pesat, berita kesehatan gizi menjadi sumber penting bagi masyarakat untuk mendapatkan informasi terbaru mengenai isu-isu gizi. Perkembangan informasi digital yang pesat telah mengubah cara mencari informasi tentang kesehatan gizi. Karena kemudahan akses informasi digital, masyarakat saat ini lebih sering mencari informasi kesehatan secara *online* (Wahyu et al., 2023). Menurut survei yang dilakukan *Reuters Institute* yang bertajuk *Digital News Report* 2023, media *online* masih mendominasi jadi sumber berita utama masyarakat Indonesia dari tahun 2021 hingga tahun 2023 walaupun trennya menurun dalam dua tahun terakhir, 89% pada tahun 2021 menurun menjadi 88% pada tahun 2022 dan menurun drastis menjadi 84% pada tahun 2023 (Annur, dalam *Reuters Institute* 2023).

Seiring dengan jumlah berita kesehatan gizi yang tersedia secara *online* terus meningkat, memberikan tantangan baru yaitu sulitnya dalam menganalisis topik utama yang menjadi isu dalam berita kesehatan gizi ini. Hal ini disebabkan bidang gizi mencakup topik yang sangat luas seperti gizi olahraga, gizi lanjut usia, gizi ibu hamil dan menyusui, dan lain sebagainya. Mengetahui topik dari suatu bacaan secara manual memerlukan banyak waktu untuk memeriksa semua bacaan pada bidang kesehatan yang ada di Indonesia (Sahria & Fudholi, 2020).

Tantangan lainnya adalah sulit untuk melacak perkembangan atau tren topik gizi dari waktu ke waktu. Misalnya, pada saat pandemi virus corona, topik terkait COVID-19 mendominasi berita kesehatan (Setiawan & Stellarosa, 2021). Namun setelah pandemi mereda, topik yang menjadi perhatian media dan masyarakat kembali beragam.

Salah satu solusi untuk mengatasi tantangan tersebut adalah dengan menerapkan *topic modeling*. *Topic modeling* merupakan dokumen teks yang terdiri kata-kata yang dituliskan dalam banyak dokumen yang dapat dinyatakan dengan kombinasi kata-kata yang saling terkait dan teknik yang dapat digunakan untuk menyimpulkan suatu topik yang tersembunyi dalam sebuah dokumen teks.

Karena *topic modeling* mewakili dari setiap dokumen sebagai kombinasi kompleks dari beberapa topik dan setiap topiknya sebagai kombinasi kompleks dari beberapa kata, kemudian juga sebagai alat *text mining* untuk mengklasifikasikan sebuah dokumen berdasarkan hasil kesimpulan topik (Nugroho & Alamsyah, 2018).

Terdapat beberapa Teknik yang dapat digunakan untuk *topic modeling*, diantaranya yaitu *Vector Space Model* (VSM), *Latent Semantic Analysis* (LSA), *Probabilistic Latent Semantic Analysis* (PLSA) dan *Latent Dirichlet Allocation* (LDA). Pada penelitian ini nantinya akan menerapkan *topic modeling* dengan menggunakan *Latent Dirichlet Allocation* (LDA) pada kumpulan berita kesehatan gizi. LDA adalah sebuah metode *topic modeling* yang digunakan untuk menentukan pola pada sebuah dokumen yang dapat menghasilkan topik (Prasanna & Rao, 2019). LDA adalah metode yang digunakan untuk menganalisis pada dokumen yang sangat besar. LDA juga digunakan untuk meringkas, melakukan pengelompokan, menghubungkan maupun memproses data. Dengan menerapkan *topic modeling* menggunakan LDA ini pada kumpulan berita kesehatan gizi, peneliti berharap dapat memperoleh topik utama beserta perkembangan tren topik gizi dari waktu ke waktu.

Salah satu penelitian yang menerapkan *topic modeling* menggunakan LDA untuk mengetahui bagaiamana tren topik dalam bidang kesehatan adalah penelitian yang dilakukan Liu et al., (2021) dengan judul *Tracing the Pace of COVID-19 Research: Topic Modeling and Evolution* yang telah berhasil mengalokasikan ke dalam 50 topik penelitian utama yang relevan dengan COVID-19 berdasarkan abstrak. Selanjutnya dari 50 topik tersebut ditemukan 5 topik paling popular yaitu Topik 37, Topik 43, Topik 33, Topik 30, dan Topik 35. Topik-topik ini tentang estimasi dan prediksi penularan COVID-19 (Topik 37), diagnosis/konfirmasi (Topik 43 dan Topik 30), dan aplikasi terkait teknologi *deep learning* (Topik 33 dan Topik 35). Penulis menemukan bahwa Topik 37 menjadi topik utama dengan periode sebagian besar dari pertengahan Januari hingga pertengahan Februari, dan Topik 43 sebagian besar dari akhir Februari hingga pertengahan Maret. Sedangkan, Topik 33, 30, dan 35 relatif stabil sepanjang seluruh periode.

Berdasarkan hal tersebut penulis mengajukan judul "Penerapan *Topic Modeling* Menggunakan *Latent Dirichlet Allocation* (LDA) Dalam Berita Kesehatan Gizi" untuk mengetahui bagaimana penerapan *topic modeling*, khususnya metode *Latent Dirichlet Allocation* (LDA), dapat membantu penemuan topik utama yang dikelompokan menjadi tren topik berita kesehatan terkait gizi.

1.2 Rumusan Masalah

Rumusan masalah dari penelitian ini adalah sebagai berikut:

1. Bagaimana penerapan *topic modeling* dalam berita kesehatan gizi mengunakan *Latent Dirichlet Allocation* (LDA)?
2. Bagaimana kinerja model *Latent Dirichlet Allocation* (LDA) dalam pengelompokan tren topik berita kesehatan gizi?

1.3 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Menganalisis penerapan *topic modeling* dalam berita kesehatan gizi menggunakan *Latent Dirichlet Allocation* (LDA).
2. Menganalisis kinerja metode *Latent Dirichlet Allocation* (LDA) dalam pengelompokan tren topik berita kesehatan gizi.

1.4 Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut:

1. Masyarakat, penelitian ini diharapkan dapat dijadikan informasi oleh masyarakat mengenai topik-topik utama dan tren topik dalam berita kesehatan gizi di Indonesia.
2. Peneliti, penelitian ini diharapkan dapat menjadi referensi bagi akademisi dan peneliti di masa depan yang tertarik untuk melakukan penelitian seputar *topic modeling* pada institusi Pendidikan.

1.5 Ruang Lingkup Penelitian

Pada penelitian ini akan dibatasi ruang lingkup pembahasannya yaitu:

1. Menggunakan metode *Latent Dirichlet Allocation* (LDA).
2. Menggunakan data berita berbahasa Indonesia dengan waktu pengambilan per tahun dari tahun 2019 sampai dengan tahun 2023.
3. Data bersumber dari beberapa portal berita *online* nasional Indonesia yang di antaranya Detik.com, CNN Indonesia, IDN Times dan Pikiran Rakyat.

BAB II
TINJAUAN PUSTAKA

2.1 Berita Kesehatan Gizi

Istilah gizi berasal dari bahasa Arab "*giza*" yang berarti zat makanan, dalam bahasa Inggris dikenal dengan istilah *nutrition* yang berarti bahan makanan atau zat gizi atau sering juga diartikan sebagai ilmu gizi. Lebih luas, gizi diartikan sebagai suatu proses organisme menggunakan makanan yang dikonsumsi secara normal melalui proses pencernaan, penyerapan, transportasi, penyimpanan, metabolisme dan pengeluaran zat gizi untuk mempertahankan kehidupan, pertumbuhan dan fungsi normal organ tubuh serta untuk menghasilkan tenaga (Irianto, 2006).

Kesehatan gizi mengacu pada dampak makanan dan komponennya terhadap pemeliharaan, perbaikan, pertumbuhan, dan fungsi tubuh secara keseluruhan, yang memainkan peran penting dalam kesejahteraan fisik, sosial, intelektual, emosional, dan spiritual (Klohe et al., 2017). Kesehatan gizi melibatkan pemahaman individu tentang nutrisi dan bagaimana mengontrol pola makan dan aktivitas fisik mereka. Gizi yang cukup sangat penting untuk menyediakan energi, meningkatkan kekebalan tubuh, dan mendukung perkembangan yang optimal, terutama pada anak di bawah usia tiga tahun, di mana praktik gizi awal secara signifikan mempengaruhi hasil kesehatan jangka panjang (Haldar et al., 2022).

Berita merupakan laporan tercepat mengenai fakta atau ide terbaru yang benar, menarik, dan penting bagi sebagian khalayak, melalui media berkala seperti surat kabar, radio, televisi, atau media *online* internet (Sumadiria, 2005). Berita kesehatan gizi mengacu pada informasi yang disebarkan melalui berbagai saluran media, seperti surat kabar, platform *online*, dan media sosial, dengan fokus pada topik yang berkaitan dengan makanan, diet, dan gizi (Ellis & Evans, 2022).

2.2 *Text Mining*

Text mining merupakan suatu teknik yang digunakan untuk menangani dari pengelompokan, klasifikasi, ekstraksi informasi dan pengambilan informasi (Hudaya et al., 2019). *Text mining* merupakan proses ketika pengguna berinteraksi dengan kumpulan dokumen dari waktu ke waktu dengan menggunakan kumpulan analisis. *Text mining* merupakan suatu analisis data yang terdapat bahasa alami dengan menggunakan teknik dan alat untuk merancang, menemukan dan mengesktrak pengetahuan pada data yang tidak terstruktur. Pada data yang tidak terstruktur sehingga dapat menjadi data dengan topik yang lebih terstruktur dengan mengubah kata atau kalimat (Kabiru & Sari, 2019).

Text mining dapat menyelesaikan masalah dengan mengadopsi dan mengembangkan banyak teknik dari bidang lain, seperti *data mining*, *information retrieval*, statistik matematik, *machine learning*, linguistik, *Natural Language Processing* (NLP), dan *visualization*. *Text mining* bertujuan untuk mengekstrak informasi yang bernilai dari berbagai sumber data. Dengan demikian, sumber data yang dimanfaatkan dalam *text mining* adalah kumpulan dokumen yang memiliki format yang tidak terstruktur, melalui proses identifikasi dan pengeksplorasian pola yang menarik (Anwar, 2022).

2.3 *Preprocessing*

Preprocessing data adalah suatu proses untuk membersihkan data dari gangguan dan mengubah menjadi bentuk yang siap untuk diolah lebih lanjut yang bertujuan untuk meningkatkan akurasi, konsistensi, dan keandalan dari sebuah dataset karena pada dasarnya dataset yang digunakan bersifat mentah dan laten. *Preprocessing* bertujuan untuk mempercepat pemroresan data karena proses ini menghilangkan beberapa bagian yang tidak dibutuhkan. Hasil yang diperoleh dari *preprocessing* akan berupa nilai numerik sehingga dapat dijadikan sebagai sumber data yang dapat diolah lebih lanjut (Socrates et al., 2016). Pada proses *preprocessing* terdiri dari *case folding*, *remove punctuation*, *number removal*, *stopword removal*, *stemming*, dan *tokenizing*.

2.3.1 *Case Folding*

Case folding adalah proses dimana akan mengubah karakter dari huruf besar (*upper case*) menjadi huruf kecil (*lower case*). Kemudian karakter yang akan diterima hanya "a" hingga "z" (Socrates et al., 2016). Dalam hal ini denan tujuan untuk dapat menghindari adanya dua kata yang sama namun dianggap berbeda oleh program yang dikarenakan perbedaan huruf kapital dan huruf kecil (Hadi et al., 2017).

2.3.2 *Punctuation Removal*

Punctuation removal merupakan proses dilakukan untuk menghapus karakter tanda baca, markup/html/tag, spesial karakter ($, %, &, stc) yang terdapat dalam dataset karena mereka tidak memberikan informasi yang signifikan pada saat analisis teks dan hanya memakan ruang penyimpanan yang tidak perlu (Cendana & Permana, 2019).

2.3.3 *Number Removal*

Number removal atau penghapusan angka bertujuan untuk menghilangkan informasi numerik yang tidak relevan, seperti tanggal, tahun, nomor urut, atau angka lainnya yang tidak berkontribusi secara substansial terhadap makna teks. Penghapusan angka melibatkan penghilangan informasi numerik yang tidak relevan dan dapat dibuang dari kumpulan data untuk meningkatkan kualitas data dan meningkatkan efisiensi model selama pemrosesan data pembelajaran mesin (Y et al., 2022).

2.3.4 *Stopword Removal*

Stopword removal merupakan proses penghilangan kata tidak penting melalui pengecekan kata-kata hasil *parsing* dalam dokumen apakah termasuk di dalam daftar kata tidak penting (*stoplist*) atau tidak. *Stopword removal* bertujuan agar data menjadi lebih bersih dan meminimalisir terjadinya *noise* pada data sehingga yang tersisa dalam *stoplist* dianggap sebagai kata-kata penting atau *keywords* (Socrates et al., 2016).

2.3.5 *Stemming*

Stemming adalah proses mendapatkan kata dasar dari kata berimbuhan, yang juga dikenal sebagai penggabungan. *Stemming* melakukan tugas ini dengan tujuan untuk mengubah kata menjadi bentuk dasar yang umum (Kaur & Buttar, 2018). *Stemming* memainkan peran penting dalam aplikasi text mining seperti kategorisasi teks, peringkasan teks, dan pencarian informasi dengan menyederhanakan ukuran kosakata dan meningkatkan efektivitas pencarian untuk memberikan hasil yang relevan (Gadri & Moussaoui, 2015).

2.3.6 *Tokenizing*

Tokenizing merupakan proses membagi suatu teks atau pemotongan *string input* dari tiap kata yang menyusunnya (Socrates et al., 2016). dengan tujuan untuk memecah kalimat menjadi kata, frasa, dan entitas penting lainnya atau yang biasa disebut sebagai token dari sebuah teks. Selain itu, agar dalam proses yang selanjutnyaa menjadi lebih mudah mengenai penghitungan kata, pembobotan kata sampai dengan transformasi kedalam bentuk vektor yang berdimensi tinggi (Hadi et al., 2017).

2.4 *Bigram* dan *Trigram*

Bigram dan *trigram* merupakan bagian dari model *N-Gram*. *Bigram* adalah sebuah *N-Gram* yang terdiri dari 2 item *sequence* sedangkan *trigram* adalah *N-Gram* yang terdiri dari 3 item *sequence*. Model *N-Gram* adalah sebuah model statistik yang dirancang untuk memprediksi item berikutnya dalam urutan item (Jurafsky & Matrin, 2023). Item yang dimaksud berupa huruf/karakter, kata, atau yang lainnya sesuai dengan aplikasi. Model ini awalnya dikembangkan oleh ahli matematika Rusia, Andrei Markov, pada awal abad ke-20. Markov menggunakan konsep *Markov chain* untuk memprediksi apakah huruf berikutnya dalam teks Pushkin's Eugene Onegin akan menjadi vokal atau konsonan. Ia mengklasifikasikan 20,000 huruf sebagai V (vokal) atau C (konsonan) dan menghitung probabilitas bigram dan trigram bahwa huruf berikutnya akan menjadi vokal diberikan huruf sebelumnya atau dua huruf sebelumnya (Jurafsky

& Matrin, 2023). Salah satu aplikasi model *N-Gram* yang berbasis kata yang digunakan untuk memprediksi kata berikutnya dalam urutan kata tertentu. Sebuah *N-Gram* hanyalah sebuah wadah kumpulan kata dengan masing-masing memiliki panjang *n* kata.

Pada pembagian karakter, *N-Gram* terdiri dari substring sepanjang n karakter dari sebuah string daalam definisi lain *N-Gram* adalah potongan sejumlah n karakter dari sebuah string. Metode *N-Gram* digunakana untuk mengambil potongan-potongan karakter huruf sejumlah n dari sebuah kata yang secara kontinuitas dibaca dari teks sumber hinnga akhir dari dokumen. Sebagai contoh kata "TEXT" dapat diuraikan ke dalam beberapa *N-Gram* berikut:

- *Uni-Gram*: T, E, X, T
- *Bi-Gram*: TE, EX, XT
- *Tri-Gram*: TEX, XET

Penggunaan *bigram* dan *trigram* untuk meminimalisir pemenggalan kata yang menghilangkan makna kata. Selain itu, keunggulan menggunakan *N-Gram* yang tidak terlalu sensitif terhadap kesalahan penulisan yang terdapat pada suatu dokumen (Hanafi, 2009).

2.5 *Term Frequency-Invers Document Frequency* (TF-IDF)

Pembobotan kata adalah proses pemberian bobot pada setiap kata yang terdapat dalam sebuah dokumen. Salah satu metode paling popular untuk mencari informasi peringkat berdasarkan frekuensi kata adalah metode TF-IDF (*Term Frequency-Inverse Document Frequency*) (Gunawan et al., 2018). TF-IDF merupakan gabungan dari dua metode yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). Pendekatan dalam pembobotan kata yang paling banyak diterapkan adalah *term frequency* (TF). Faktor ini menyatakan banyaknya kemunculan suatu kata dalam suatu dokumen. Semakin sering suatu kata muncul dalam sebuah dokumen, berarti semakin penting kata tersebut (Siregar et al., 2017).

Term Frequency (TF) memiliki beberapa jenis algoritma sebagai berikut (Yoren, 2018):

1. *Binary Term Frequency* (TF Binary), yaitu penyeragaman bobot pada dokumen yang terdiri dari 0 jika *term* tidak muncul dan 1 jika *term* muncul
2. *Raw Term Frequency* (TF Murni), yaitu nilai TF yang didapatkan berdasarkan frekuensi dari suatu *term* pada dokumen. Jika muncul *term* sebanyak 3 kali, maka *term* tersebut akan bernilai 3
3. *Logarithmic Term Frequency* (TF Logaritmik), yaitu nilai TF dengan fungsi logaritmik untuk perhitungannya menghindari dokumen yang mengandung sedikit *term*, namun memiliki frekuensi yang tinggi

$$TF = 1 + \log(TF) \tag{1}$$

4. *Augmented Term Frequency* (TF Normalisasi), yaitu perhitungan nilai TF menggunakan perbandingan antara frekuensi sebuah kata dengan jumlah kesuluruhan kata pada dokumen

$$TF = 0.5 + 0.5\left(\frac{TF}{maxTF}\right) \tag{2}$$

Dimana nilai *max TF* merupakan jumlah muncul kata (t) terbanyak pada dokumen yang sama.

Inverse Document Frequency (IDF) menunjukkan hubungan ketersediaan sebuah *term* dalam seluruh dokumen. Semakin rendah nilai TF, maka akan semakun tinggi nilai IDF (Utomo et al., 2021). IDF digunakan untuk menentukan pentingnya kata dalam dokumen, kata-kata yang muncul dalam sejumlah besar dokumen akan menjadi lebih penting atau memiliki bobot yang lebih besar. Berikut persamaan untuk menghitung IDF.

$$IDF_{(t)} = ln\left(\frac{D}{DF_{(t)}}\right) \tag{3}$$

dimana,

$IDF_{(t)}$ = nilai *Inverse Document Frequency* suatu *term* (t)

D = jumlah keseluruhan dokumen

$DF_{(t)}$ = jumlah dokumen yang mengandung *term* (t)

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan algoritma yang digunakan untuk menghitung bobot setiap kata pada masing-masing dokumen terhadap kata kunci. *Term Frequency* (TF) merupakan cara untuk mencari bobot dari sebuah dokumen. Semakin banyak jumlah kemunculan sebuah kata maka akan mempengaruhi besar bobotnya, sedangkan *Inverse Document Frequency* (IDF) merupakan proses untuk mneghitung penyebaran kata dalam dokumen. Penyebaran kata yang tidak sesuai bisa mempengaruhi hasil dari perhitungan bobot pada dokumen (Rizkia, 2019). Semakin banyak kata yang muncul pada dokumen maka kepentingan kata tersebut menjadi sedikit sehingga bobot yang diberikan pun kecil begitupun sebaliknya apabila kemunculan kata semakin sedikit maka bobot yang diberikan menjadi besar sehingga kata tersebut menjadi penting (Fauzi et al., 2019). *Document Frequency* (DF) adalah jumlah dokumen dimana kata-kata tertentu muncul. Nilai pembobotan TF-IDF akan tinggi jika nilai TF besar dan kata yang diamati tidak ditemukan dalam banyak dokumen (Laxmi et al., 2018). Oleh karena itu, untuk mencari nilai TF-IDF menggunakan persamaan 4.

$$w_{(t,d)} = TF_{t,d} \times IDF_t \tag{4}$$

Jika frekuensi *term* sama dengan jumlah dokumen, maka hasil perhitungan $IDF_t = 0$. Untuk menghindari hasil $w_{(t,d)} = 0$, dapat ditambahkan nilai 1 pada hasil perhitungan IDF (Apriani et al., 2021), seperti pada persamaan 5.

$$w_{(t,d)} = TF_{t,d} \times (IDF_t + 1) \tag{5}$$

dimana:

$w_{(t,d)}$ = bobot dari suatu kata (t) dalam satu dokumen

$TF_{t,d}$ = frekuensi kemuculan suatu kata (t) dalam dokumen (d)

IDF_t = nilai *Inverse Document Frequency* suatu *term* (t)

2.6 *Topic Modeling*

Konsep *topic modeling* menurut Blei (2003) terdiri dari entitas-entitas yaitu "kata", "dokumen", dan "*corpora*". "Kata" dianggap sebagai unit dasar dari data diskrit dalam dokumen, didefinisikan sebagai item dari kosa kata yang diberi indeks untuk setiap kata unik pada dokumen. "Dokumen" adalah susunan N kata-kata. Sebuah *corpus* adalah kumpulan M dokumen dan *corpora* merupakan bentuk jamak dari *corpus*. Sementara "*topic*" adalah distribusi dari beberapa kosakata yang bersifat tetap. Secara sederhana, setiap dokumen dalam *corpus* mengandung proporsi tersendiri dari topik-topik yang dibahas seusai kata-kata yang terkandung di dalamnya (Blei, 2003).

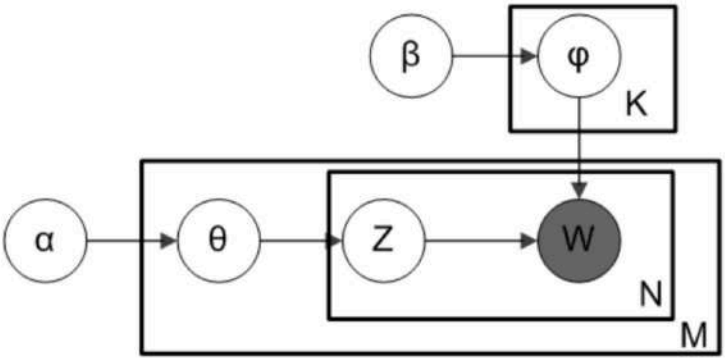
Ide dasar dari *topic modeling* adalah bahwa sebuah topik terdiri dari kata-kata tertentu yang menyusun topik tersebut, dan dalam satu kemungkinan terdiri dari beberapa topik dengan probabilitas masing-masing. Namun secara pemahaman manusia, dokumen-dokumen merupakan objek yang diamati, sedangkan topik, distribusi topik per-dokumen, dan penggolongan setiap kata pada topik per-dokumen merupakan struktur tersembunyi, maka dari itu *topic modeling* bertujuan untuk menemukan topik dan kata-kata yang terdapat pada topik tersebut (Blei, 2012). Untuk menemukan topik yang berada antara teks dengan *topic modeling* menganalisis dari teks asli, bagaimana topik-topik dapat saling terhubung satu dengan yang lain, bagaimana tema-temabisa berubah dari waktu ke waktu, sehingga bisa dikembangkan untuk pencarian, atau dengan meringkas teks yang terdapat dalam dokumen teks (Putra & Kusumawardani, 2017).

Topic modeling merupakan dokumen teks yang terdiri kata-kata yang dituliskan dalam banyak dokumen yang dapat dinyatakan dengan kombinasi kata-kata yang saling terkait dan teknik yang dapat digunakan untuk menyimpulkan suatu topik yang tersembunyi dalam sebuah dokumen teks. Karena *topic modeling* mewakili dari setiap dokumen sebagai kombinasi kompleks dari beberapa topik dan setiap topiknya sebagai kombinasi kompleks dari beberapa kata, kemudian juga sebagai alat *text mining* untuk mengklasifikasikan sebuah dokumen berdasarkan hasil kesimpulan topik (Nugroho & Alamsyah, 2018).

2.7 *Latent Dirichlet Allocation* (LDA)

Latent Dirichlet Allocation merupakan salah satu metode yang dipilih dalam melakukan analisis pada dokumen yang sangat besar. LDA sendiri dapat digunakan meringkas, melakukan klasterisasi, menghubungkan maupun memproses data yang sangat besar dikarenakan LDA menghasilkan daftar topik yang diberi bobot untuk masing-masing dokumen. Distribusi yang digunakan untuk mendapatkan distribusi topik untuk setiap dokumen disebut distribusi *Dirichlet*, kemudian dalam proses generatif untuk LDA, hasil yang didapatkan dari *Dirichlet* digunakan untuk mengalokasikan kata-kata pada dokumen untuk topik yang berbeda. *Latent* didefinisikan sebagai sesuatu yang ada tetapi tidak terlihat. Dokumen-dokumen dalam LDA merupakan objek yang diamati, sedangkan topik, distribusi topik per-dokumen, penggolongan setiap kata pada topik per-dokumen merupakan struktur tersembunyi (Blei, 2012).

Menurut Natalia et al., (2021), dalam penerapannya LDA membentuk sebuah sistem untuk menetukan beberapa kelompok berdasarkan data atau informasinya yang terdiri dari kata-kata. Metode LDA meupakan improvisasi metode dari dua metode yang sebelumnya telah diperkenalkan terlebih dahulu yaitu metode *Probabilistic Latent Semantic Analysis* (PLSA) dan *Latent Semantic Analysis* (LSA).



Gambar 1. *Plate notation* LDA
Sumber: (Setijohatmo et al., 2020)

Berikut merupakan formulasi dari *Latent Dirichlet Allocation*:

$$p(w, z, \theta, \varphi|\alpha, \beta) = p(\varphi|\beta)p(\theta|\alpha)p(z|\theta)p(w|\varphi_k) \quad (6)$$

Dari Gambar 1 di atas dapat diketahui bahwa metode LDA memiliki beberapa variabel yaitu:

α = distribusi topik dalam dokumen pada suatu corpus

θ = probabilitas *i* dokumen yang mengandung *j* topik

β = probabilitas *i* topik yang mengandung *j* kata

M = jumlah dokoumen

N = jumlah kata dalam setiap dokumen

w = kata dalam dokumen

z = topik yang terbentuk

φ = distribusi kata terhadap topik dalam corpus

Menurut Prihatini et al., (2017) menyatakan bahwa algoritma LDA terdiri dari proses inisialisasi, proses berulang dan *sampling*, dan proses membaca parameter akhir (Syaifuddin et al., 2020).

1. Tahap inisialisasi merupakan proses penentuan frekuensi kemunculan kata pada setiap *file* teks. Proses ini dilakukan pada teks hasil *preprocessing* data. Proses inisialisasi dilakukan dengan langkah:
    a. Menentukan indeks dari setiap kata dalam dokumen
    b. Menghitung frekuensi kemunculan setiap kata pada setiap dokumen menggunakan *Bag-of-Words* (BoW)
    c. Menentukan topik setiap kata dengan random berdasarkan nilai frekuensi kemunculan kata (*z0*), LDA membutuhkan nilai topik yang ditentukan terlebih dahulu selain itu, dalam algoritma LDA, tidak ada nilai awal yang diberikan untuk setiap kata dalam dokumen, sehingga setiap kata memiliki nilai ketidakpastian
    d. Menentukan matriks kata-topik dan dokumen-topik
    e. Menghitung jumlah total dari distribusi kata-topik dan dookumen-topik, dan menyimpan hasil matriks. Distribusi kata, *Wi*, pada tiap topik, *Zi*, *NW* dilihat pada persamaan berikut

$$NW = \begin{Bmatrix} NW_{w1,z1} & NW_{w1,z2} & \cdots & NW_{w1,zk} \\ NW_{w2,z1} & NW_{w2,z2} & \cdots & NW_{w2,z1} \\ \vdots & \vdots & \ddots & \vdots \\ NW_{wn,z1} & NW_{wn,z2} & \cdots & NW_{wn,zk} \end{Bmatrix} \quad (7)$$

dimana,

w_n = *term* ke-*n* pada vocab

z_k = topik ke-*k*

$NW_{wn,zk}$ = banyak *term* ke-*n* yang berlabel topik ke-*k*

Kemudian membuat matriks distribusi topik, *Zk*, pada tiap dokumen:

$$ND = \begin{Bmatrix} ND_{d1,z1} & ND_{d1,z2} & \cdots & ND_{d1,zk} \\ ND_{d2,z1} & ND_{d2,z2} & \cdots & ND_{d2,z1} \\ \vdots & \vdots & \ddots & \vdots \\ ND_{dn,z1} & ND_{dn,z2} & \cdots & ND_{dn,zk} \end{Bmatrix} \quad (8)$$

dimana,

d_m = dokumen ke-*m*

z_k = topik ke-*k*

$ND_{dn,zk}$ = banyak label topik ke-*k* pada dokumen ke-*n*

Pada akhir proses inisialisasi, dilakukan proses perhitungan jumlah setiap topik dalam dokumen (*ND*) dan jumlah setiap kata dalam topik (*NW*) yang akan digunakan dalam proses iterasi dan sampling topik. Nilai total semua *ND* juga dihitung sebagai SUMND dan nilai total semua *NW* sebagai SUMNW. Nilai dari SUMND dan SUMNW digunakan untuk mengurangi dan menambah nilai *ND* dan *NW* di sembarang perubahan topik yang terjadi pada setiap kata. Jumlah distribusi *NW* dan *ND* sebagai berikut:

$$NWSum_{zi} = \sum_{j=1}^{m} NW_{wj,zi} \ldots\ldots\ldots \quad (9)$$

$$NDSum_{di} = \sum_{j=1}^{k} ND_{wi,zj} \ldots\ldots\ldots \quad (10)$$

dimana,

$NWSum_{zi}$ = jumlah seluruh kata dalam setiap topik

$NDSum_{di}$ = jumlah seluruh topik dalam setiap dokumen

$NW_{wj,zi}$ = jumlah setiap kata dalam topik

$ND_{wi,zj}$ = jumlah topik dalam dokumen

2. Tahap *sampling* topik merupakan proses penentuan tpoik baru dari setiap kata pada setiap dokumen. Proses ini dilakukan pada teks hasil *preprocessing*. Proses *sampling* topik dilakukan dengan langkah:

   a. Menghitung probabilitas kata pada topik

$$\emptyset_{ij} = \frac{C_{ij}^{WT} + \beta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\beta} \tag{11}$$

dimana,

$\emptyset_{ij}$ = probabilitas dari kata *i* untuk topik *j*

C_{ij}^{WT} = jumlah kata *i* pada topik *j*

WT = kata-topik

β = parameter sebaran *Dirichlet* distribusi kata terhadap topik

$\sum_{k=1}^{W} C_{kj}^{WT}$ = jumlah seluruh kata pada topik *j*

k = indeks topik

W = jumlah seluruh kata pada dokumen

Menghitung probabilitas dokumen topik:

$$\theta_{ij} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha} \tag{12}$$

dimana,

θ_{dj} = proporsi topik *j* dalam dokumen *d*

C_{dj}^{DT} = jumlah topik *j* pada dokumen *d*

DT = dokumen-topik

α = parameter sebaran *Dirichlet* distribusi topik terhadap dokumen

$\sum_{k=1}^{T} C_{dk}^{DT}$ = jumlah seluruh topik pada dokumen *d*

T = jumlah seluruh topik yang sudah ditemukan

b. Menetukan topik baru dari setiap kata dengan distribusi multinomial (posterior) berdasarkan nilai probabilitas kata tertinggi. Parameter distribusi posterior dimana bobot tersebut diperoleh dari nilai distribusi probabilitas kata-topik dikali distribusi probabilitas topik-dokumen.

$$P(z_i = j|z_i, w_i, d_i) = \frac{C_{ij}^{WT} + \beta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\beta} \times \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha} \quad (13)$$

c. Menyimpan hasil distribusi posterior

d. Langkah-langkah ini dilakukan sebanyak *n* perulangan sampai mencapai kondisi konvergen.

3. Tahap perhitungan parameter final merupakan proses untuk menghitung jumlah dokumen per topik dan jumlah kata per topik berdasarkan matriks kata-topik dan dokumen-topik yang telah konvergen.

2.8 *Gibbs Sampling*

Latent Dirichlet Allocation (LDA) adalah model probabilistik yang digunakan untuk mengidentifikasi topik tersembunyi dalam sebuah koleksi dokumen. LDA menganggap setiap dokumen sebagai kombinasi dari beberapa topik, di mana setiap topik direpresentasikan oleh pola distribusi kata-kata yang mungkin muncul dalam topik tersebut (Blei, 2003). Salah satu algoritma yang paling umum digunakan untuk memperkirakan parameter model LDA adalah *Gibbs sampling*.

Gibbs sampling adalah teknik *Markov Chain Monte Carlo* (MCMC) yang digunakan untuk menghitung distribusi probabilitas posterior pada model LDA. Algoritma ini melakukan sampling berulang dari distribusi posterior bersyarat untuk setiap penugasan topik kata, dengan mengasumsikan penugasan topik untuk kata-kata lain telah diketahui (Griffiths & Steyvers, 2004). Proses ini diulang sampai konvergen menuju distribusi stasioner. Pada implementasi *Gibbs sampling* untuk LDA, setiap kata dalam *corpus* ditetapkan ke salah satu topik laten

berdasarkan probabilitas bersyarat. Probabilitas bersyarat ini bergantung pada frekuensi kata dalam topik dan frekuensi topik dalam dokumen (Griffiths & Steyvers, 2002). Setelah iterasi cukup, distribusi kata-kata untuk setiap topik dan distribusi topik untuk setiap dokumen dapat diperkirakan dari penugasan topik kata.

Gibbs sampling dalam konteks *Latent Dirichlet Allocation* (LDA) memiliki formulasi tersendiri yang digunakan untuk memperbarui distribusi topik bagi setiap kata dalam dokumen. Formulasi ini didasarkan pada distribusi posterior dari topik untuk setiap kata, dengan mempertimbangkan distribusi kata dalam topik dan distribusi topik dalam dokumen. Formulasi *Gibbs sampling* untuk LDA ditunjukkan pada persamaan 14 (Griffiths & Steyvers, 2004).

Untuk setiap kata w dalam dokumen d:

1. Hapus penugasan topik saat ini untuk kata w.
2. Hitung probabilitas kondisional untuk setipa topik k:

$$P(z_{di} = k | z_{-di}, w, \alpha, \beta) \propto \frac{(n_{dk}^{-di} + \alpha_k)(n_{kv}^{-di} + \beta)}{n_k^{-di} + V\beta} \quad (14)$$

dimana,

$P(z_{di} = k | z_{-di}, w, \alpha, \beta)$ = probabilitas kata w di posisi i dalam dokumen d diberi topik k

n_{dk}^{-di} = jumlah kata dalam dokumen d yang diberi topik k, tidak termasuk kata w di posisi i

n_{kv}^{-di} = jumlah kemunculan kata v dalam topik k, tidak termasuk kata w di posisi i

n_k^{-di} = total jumlah kata yang diberi topik k dalam semua dokumen, tidak termasuk kata w di posisi i

α_k dan β = hyperparameter *Dirichlet* untuk distribusi topik dalam dokumen dan distribusi kata dalam topik

V = jumlah total kata unik dalam *corpus*

3. Pilih topik baru untuk kata w berdasarkan distribusi probabilitas yang telah dihitung.

2.9 *Topic Coherence*

Topic coherence digunakan untuk mengevalusai model topik. *Topic coherence* yaitu dimana satu set dari kata-kata yang dihasilkan pada model topik dengan nilai berdasarkan tingkat koherensi atau dalam diinterpretasi oleh manusia dengan tingkat kemudahannya. *Topic coherence* didefinisikan sebagai rata-rata dari skor kemiripan kata berpasangan dari kata-kata dalam topik. Setiap topik dalam model terdiri dari kata-kata dan *topic coherence* diterapkan pada *N* kata teratas (Gangadharan & Gupta, 2020).

Coherence score adalah metrik kinerja yang sering digunakan untuk menilai teknik pemodelan topik. Nilai *coherence* memberikan kisaran yang realistis untuk mengidentifikasi jumlah topik dalam dokumen. Setiap topik yang dihasilkan memiliki daftar kata yang mirip dengan kelompok kata. Kisaran ini menyajikan skor kemiripan rata-rata dari kata-kata yang berhubungan dengan topik. Model yang baik adalah model yang menghasilkan topik dengan nilai *coherence* yang tinggi (George & Srividhya, 2020).

Nilai *coherence* yang baik dapat bervariasi, beberapa penelitian menunjukkan bahwa nilai rata-rata nilai *coherence* yang baik untuk model LDA berkisar antara 0,4 hingga 0,7. Penelitian yang dilakukan oleh M C et al., (2022) dengan judul *Comparative Analysis of Research Papers Categorization using LDA and NMF Approaches* menemukan bahwa LDA memiliki nilai *coherence* rata-rata sebesar 0,5282. Penelitian lain yang dilakukan oleh Alash & Al-Sultany, (2020) dengan judul *Improve topic modeling algorithms based on Twitter hashtags* mencatat bahwa dengan konfigurasi optimal, LDA dapat mencapai nilai *coherence* hingga 0,6047. Penelitian yang dilakukan oleh Robert et al., (2021) dengan judul *Investigating Individuals' Perceptions Regarding the Context Around the Low Back Pain Experience: Topic Modeling Analysis of Twitter Data* mendapatkan bahwa LDA adalah model yang terbaik dari algoritma lainnya dengan nilai *coherence* 0,562. Penelitian lainnya oleh Fahlevvi & Sn, (2022) dengan judul *Topic Modeling on Online News.Portal Using Latent Dirichlet Allocation (LDA)*

didapatkan nilai terbaik pemodelan topik menggunakan *coherence* dengan nilai 0,53. Penelitian oleh Hardiyanti et al., (2023) dengan judul *Identify Reviews of Pedulilindungi Applications using Topic Modeling with Latent Dirichlet Allocation Method* didapatkan hasil penentuan jumlah topik optimal dengan nilai *coherence* tertinggi yaitu 0,47.

Röder et al., (2015) memperkenalkan metode *Coherence Value* (CV) yang terdiri empat tahapan yaitu *segmentation*, *probability estimation*, *confirmation measure* dan *aggregation*.

1. *Segmentation*

   Segmentation membagi kata-kata dalam topik menjadi pasangan-pasangan kata. Misalnya, untuk topik T yang berisi kata-kata $w_1, w_2, \dots, w_N$, pasangan kata yang terbentuk adalah:

   $$Pairs(T) = \{(w_i, w_j) \mid 1 \leq i \leq j \leq N\} \quad (15)$$

   dimana,

   (w_i, w_j) = pasangan kata w_i dan w_j

2. *Probability Estimation*

   Tahap ini melibatkan penghitungan probabilitas kemunculan kata dan pasangan kata dalam *corpus*. Probabilitas kata w_i dan probabilitas bersama w_i dan w_j dihitung sebagai berikut:

   $$P(w_i) = \frac{Count(w_i)}{\sum_k Count(w_k)} \quad (16)$$

   $$P(w_i, w_j) = \frac{Count(w_i, w_j)}{D} \quad (17)$$

   dimana,

   $Count(w_i)$ = frekuensi kemunculan kata w_i dalam *corpus*

   $Count(w_i, w_j)$ = frekuensi kemunculan pasangan kata w_i dan w_j dalam dokumen yang sama

   D = total jumlah dokumen

3. *Confirmation Measure*

   Confirmation Measure, seperti *Pointwise Mutual Information* (PMI), digunakan untuk mengukur seberapa sering pasangan kata muncul bersama dibandingkan dengan kemunculan mereka secara individual. Formulasinya adalah:

$$PMI(w_i, w_j) = \log \frac{P(w_i, w_j) + \epsilon}{P(w_i).P(w_j)} \quad (18)$$

   dimana,

   ϵ = nilai *smoothing* yang sangat kecil untuk mneghindari pembagian dengan nol

4. *Aggregation*

   Pada tahap ini, nilai *confirmation measure* untuk semua pasangan kata dalam topik digabungkan untuk mendapatkan nilai *coherence* untuk topik tersebut:

$$C(T) = \frac{2}{|W|(|W|-1)} \sum_{1 \le i \le j \le |W|} PMI(w_i, w_j) \quad (19)$$

   dimana,

   $|W|$ = jumlah kata dalam topik T

2.10 *pyLDAvis*

PyLDAvis merupakan library yang terdapat pada python untuk visualisasi model topik yang interaktif. *PyLDAvis* dirancang untuk dapat membantu pengguna menafsirkan topik dalam model topik yang sesuai dengan sekumpulan data teks. *Packages* ini juga mengesktraksi informasi dari model topik LDA yang dipasang untuk dapat menginformasikan visualisasi berbasis web interaktif. Visualisasi yang dimaksudkan yaitu untuk digunakan dalam *notebook* IPython namun dapat juga disimpan dalam file HTML yang berdiri sendiri untuk memudahkan berbagi (Pypi.org, 2023).

2.11 *Principal Component Analysis* (PCA)

Principal Component Analysis (PCA) merupakan suatu teknik yang digunakan untuk menyederhanakan data dengan cara mengubah data secara linier untuk membentuk sistem koordinat baru dengan varians maksimum (Utami et al., 2017).

Menurut Smith dalam jurnal Rahmawati, 2020 tujuan dari prosedur PCA adalah menyederhanakan variabel yang diamati dengan cara mereduksi dimensinya. Berikut merupakan tahapan mereduksi dimensi dalam PCA:

1. Menghilangkan korelasi diantara variabel *independent* dengan melakukan transformasi variabel *independent* asal menjadi variabel baru yang tidak berkorelasi sama sekali atau sering disebut dengan *principal component*.
2. Menghapus *eigenvector* serta *eigenvalue* yang sangat kecil sehingga dimensi akan berkurang dan tidak akan membuat kehilangan data yang penting. Hal tersebut dikarenakan, *eigenvector* dengan *eigenvalue* yang besar berperan paling penting dalam proses transformasi.

Dalam konteks analisis topik, khususnya menggunakan *Latent Dirichlet Allocation* (LDA), visualisasi topik dapat menjadi tantangan karena banyaknya dimensi data yang dihasilkan oleh model LDA. *PyLDAvis* adalah alat yang populer digunakan untuk memvisualisasikan model LDA dengan tujuan mempermudah interpretasi topik yang dihasilkan. Salah satu metode yang digunakan dalam *pyLDAvis* untuk mengurangi dimensi data dan memvisualisasikannya adalah *Principal Component Analysis* (PCA) (Sievert & Shirley, 2014).

PyLDAvis menggunakan PCA untuk mereduksi matriks distribusi topik (*topic-term distribution*) dan matriks distribusi dokumen-topik (*doc-topic distribution*) menjadi dua dimensi utama. Hal ini memungkinkan visualisasi dalam bentuk plot interaktif yang menampilkan topik sebagai titik di ruang dua dimensi, dengan jarak antara titik-titik ini mencerminkan kesamaan antara topik. Dengan kata lain, topik yang serupa akan dikelompokkan bersama, sementara topik yang berbeda akan terpisah lebih jauh (Chuang et al., 2012).

2.12 *Multidimensional Scaling* (MDS)

Multidimensional Scaling (MDS) adalah teknik statistik yang digunakan untuk memvisualisasikan tingkat kesamaan atau perbedaan antara set objek. Dalam konteks *pyLDAvis*, MDS digunakan untuk memvisualisasikan hubungan antar topik yang dihasilkan dari model *Latent Dirichlet Allocation* (LDA) dalam ruang dua dimensi (Sievert & Shirley, 2014).

MDS bekerja dengan mengambil matriks jarak atau ketidaksamaan antara pasangan objek dan mencoba menemukan konfigurasi titik-titik dalam ruang dimensi yang lebih rendah (biasanya 2D atau 3D) sedemikian rupa sehingga jarak antar titik dalam ruang baru ini mendekati jarak asli semaksimal mungkin (Borg & Groenen, 2005).

Dalam konteks visualisasi topik menggunakan *pyLDAvis*, jarak antara distribusi topik sering dihitung menggunakan *Jensen-Shannon Divergence* (JSD), yang kemudian digunakan sebagai input untuk MDS. JSD adalah versi simetris dan lebih stabil dari *Kullback-Leibler* (KL) *Divergence*, yang mendefinisikan seberapa berbeda dua distribusi probabilitas (Lin, 1991). Formula untuk JSD antara dua distribusi P dan Q dapat dilihat pada persamaan 20.

$$JSD(P\|Q) = \frac{1}{2} D_{KL}(P\|M) + \frac{1}{2} D_{KL}(Q\|M) \tag{20}$$

dimana,

M $= \frac{1}{2}(P\|Q)$

D_{KL} $=$ *Kullback-Leibler Divergence*

$$D_{KL}(P\|Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right) \tag{21}$$

PyLDAvis menggunakan hasil MDS ini untuk memposisikan lingkaran-lingkaran yang merepresentasikan topik dalam visualisasi 2D. Jarak antar lingkaran mencerminkan perbedaan antar topik, sementara ukuran lingkaran menunjukkan prevalensi topik dalam *corpus* (Sievert & Shirley, 2014).