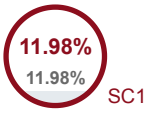
Similarity Report

Metadata

Name of the organization
Universitas Muhammadiyah Sidoarjo
Title
Karya Tulis Ilmiah Mahasiswa UMSIDA Muhammad Lazuardi Imani 181080200289
Author(s) Coordinator
perpustakaan umsidairta
Organizational unit
Perpustakaan

Record of similarities

SCs indicate the percentage of the number of words found in other texts compared to the total number of words in the analysed document. Please note that high coefficient values do not automatically mean plagiarism. The report must be analyzed by an authorized person.



4414	33450
Length in words	Length in characters

Alerts

In this section, you can find information regarding text modifications that may aim at temper with the analysis results. Invisible to the person evaluating the content of the document on a printout or in a file, they influence the phrases compared during text analysis (by causing intended misspellings) to conceal borrowings as well as to falsify values in the Similarity Report. It should be assessed whether the modifications are intentional or not.

Characters from another alphabet	R	0
Spreads	A→	0
Micro spaces	·	1
Hidden characters	R	0
Paraphrases (SmartMarks)	a	39

Active lists of similarities

This list of sources below contains sources from various databases. The color of the text indicates in which source it was found. These sources and Similarity Coefficient values do not reflect direct plagiarism. It is necessary to open each source, analyze the content and correctness of the source crediting.

The 10 longest fragments

Color of the text

NO	TITLE OR SOURCE URL (DATABASE)	NUMBER OF IDENTICAL WORDS (FRAGMENTS)
1	https://jurnal.stkippgritulungagung.ac.id/index.php/jipi/article/view/4927	33 0.75 %
2	ANALISIS SENTIMEN DAN PEMODELAN TOPIK TERHADAP PEMBANGUNAN KERETA CEPAT JAKARTA-BANDUNG MENGGUNAKAN NAIVE BAYES DAN LATENT DIRICHLET ALLOCATION (LDA) Budi Soesilo, Mulaab Mulaab Mulaab,Fatah Doni Abdul, Kamil Fajrul Ihsan;	29 0.66 %
3	https://scholarhub.ui.ac.id/jkmi/vol14/iss1/3/	28 0.63 %

4	TOPIC MODELING IN COVID-19 VACCINATION REFUSAL CASES USING LATENT DIRICHLET ALLOCATION AND LATENT SEMANTIC ANALYSIS Ulfah Malihatin S, Yulian Findawati, Uce Indahyanti;	24 0.54 %
5	TOPIC MODELING IN COVID-19 VACCINATION REFUSAL CASES USING LATENT DIRICHLET ALLOCATION AND LATENT SEMANTIC ANALYSIS Ulfah Malihatin S, Yulian Findawati, Uce Indahyanti;	23 0.52 %
6	https://archive.umsida.ac.id/index.php/archive/preprint/download/4670/33674/38002	22 0.50 %
7	Analisa Pemodelan Topik Berita Daring Menggunakan Semi-Supervised dan Fully Unsupervised Latent Dirichlet Allocation Nurdin Khoirunnisa Fi, Sutanto Taufik Edy, Ary Santoso;	21 0.48 %
8	Analisa Pemodelan Topik Berita Daring Menggunakan Semi-Supervised dan Fully Unsupervised Latent Dirichlet Allocation Nurdin Khoirunnisa Fi, Sutanto Taufik Edy, Ary Santoso;	19 0.43 %
9	https://repository.ummat.ac.id/6643/9/BAB%20V-LAMPIRAN.pdf	17 0.39 %
10	Augmenting sentiment prediction capabilities for code-mixed tweets with multilingual transformers Hashmi, Ehtesham, Shaikh, Sarang, Yayilgan, Sule Yildirim;	17 0.39 %

from RefBooks database (6.71 %)

NO	TITLE	NUMBER OF IDENTICAL WORDS (FRAGMENTS)
Source: Paperity		
1	TOPIC MODELING IN COVID-19 VACCINATION REFUSAL CASES USING LATENT DIRICHLET ALLOCATION AND LATENT SEMANTIC ANALYSIS Ulfah Malihatin S, Yulian Findawati, Uce Indahyanti;	171 (14) 3.87 %
2	Analisa Pemodelan Topik Berita Daring Menggunakan Semi-Supervised dan Fully Unsupervised Latent Dirichlet Allocation Nurdin Khoirunnisa Fi, Sutanto Taufik Edy, Ary Santoso;	47 (3) 1.06 %
3	ANALISIS SENTIMEN DAN PEMODELAN TOPIK TERHADAP PEMBANGUNAN KERETA CEPAT JAKARTA-BANDUNG MENGGUNAKAN NAIVE BAYES DAN LATENT DIRICHLET ALLOCATION (LDA) Budi Soesilo, Mulaab Mulaab Mulaab,Fatah Doni Abdul, Kamil Fajrul Ihsan;	**29 (1) 0.66 %**
4	Augmenting sentiment prediction capabilities for code-mixed tweets with multilingual transformers Hashmi, Ehtesham, Shaikh, Sarang, Yayilgan, Sule Yildirim;	17 (1) 0.39 %
5	Mekanisme Penyelesaian Sengketa Pemilihan Umum di Indonesia M. Ikhwan,Erick Benni;	16 (2) 0.36 %
6	Analisa Tingkat Pemanfaatan Fasilitas Pokok dan Fasilitas Penunjang di Pelabuhan Perikanan (PPI) Binuangeun, Kabupaten Lebak Zepanya Fernando, Nurul Muharani,Rahmawati Anisa Yuli, Ma'ruf Ma'ruf, Musonnif Miftahul;	11 (1) 0.25 %
7	PEMODELAN PERSEPSI PEMBELAJARAN ONLINE MENGGUNAKAN LATENT DIRICHLET ALLOCATION Fernanda Jerhi Wahyu;	5 (1) 0.11 %

from the home database (0.00 %)

NO	TITLE	NUMBER OF IDENTICAL WORDS (FRAGMENTS)

from the Database Exchange Program (0.00 %)

NO	TITLE	NUMBER OF IDENTICAL WORDS (FRAGMENTS)

from the Internet (5.28 %)

NO	SOURCE URL	NUMBER OF IDENTICAL WORDS (FRAGMENTS)
1	https://tangan-digi.dinamika.ac.id/uploads/proposal/21410100005-proposal-ta-2024110009.pdf	41 (4) 0.93 %
2	**https://jurnal.stkippgritulungagung.ac.id/index.php/jipi/article/view/4927**	**33 (1) 0.75 %**
3	**https://scholarhub.ui.ac.id/jkmi/vol14/iss1/3/**	**28 (1) 0.63 %**
4	https://archive.umsida.ac.id/index.php/archive/preprint/download/4670/33674/38002	22 (1) 0.50 %
5	https://repository.dinamika.ac.id/id/eprint/8025/1/21410100005-2025-UNIVERSITASDINAMIKA.pdf	17 (3) 0.39 %
6	https://repository.ummat.ac.id/6643/9/BAB%20V-LAMPIRAN.pdf	17 (1) 0.39 %
7	https://ceur-ws.org/Vol-3688/paper1.pdf	17 (1) 0.39 %
8	https://glorespublication.org/index.php/globalistik/article/download/120/58	15 (1) 0.34 %
9	https://eprints.ums.ac.id/23995/12/NASKAH_PUBLIKASI.pdf	14 (2) 0.32 %
10	https://www.warse.org/IJATCSE/static/pdf/file/ijatcse249942020.pdf	12 (1) 0.27 %
11	http://repository.ub.ac.id/168869/1/Ivan.pdf	11 (1) 0.25 %
12	https://digilib.uin-suka.ac.id/id/eprint/42685/1/16650053_BAB-I_IV-atau-V_DAFTAR-PUSTAKA.pdf	6 (1) 0.14 %

List of accepted fragments (no accepted fragments)

NO	CONTENTS	NUMBER OF IDENTICAL WORDS (FRAGMENTS)

Topic Modeling in 2019 Indonesian Election News Titles **Using the Latent Dirichlet Allocation (LDA) Method**
[Pemodelan Topik dalam Judul Berita Pemilu 2019 Indonesia **Menggunakan Metode Latent Dirichlet Allocation (LDA)]**

Muhammad Lazuardi Imani 1), Arif Senja Fitrani *, 2) **1)Program** Studi Teknik Informatika, Universitas **Muhammadiyah Sidoarjo, Indonesia 2) Program** Studi Teknik Informatika, Universitas **Muhammadiyah Sidoarjo, Indonesia *Email Penulis** Korespondensi: asfjim@ umsida.ac.id

Abstract. This research examines the reporting of the 2019 General Election in Indonesia, focusing on analysis using the Latent Dirichlet Allocation (LDA) method. With comprehensive data processing and text analysis, the study successfully identifies key topics in the media narrative related to the election. These topics include the dynamics of the general election, such as campaign strategies and the influence of significant political figures, as well as logistical and administrative elements of the election. The LDA method, enhanced with interactive visualization using pyLDAvis, allows for detailed analysis of how the media approaches these issues. The results of this research provide deep insights into how various aspects of the election are perceived and presented to the public through the media. These findings are highly relevant not only for understanding the broader context of the election but also in the context of designing effective political communication strategies for the future. This research also highlights the importance of understanding media narratives in portraying the political landscape, as well as facilitating more informed discussions regarding public policy and political participation in Indonesia.
Keywords - Latent Dirichlet Allocation (LDA), 2019 General Election, Topic Modeling

Abstrak. Penelitian ini mengkaji pemberitaan Pemilu 2019 di Indonesia dengan fokus pada analisis menggunakan metode Latent Dirichlet Allocation (LDA). Dengan pengolahan data yang komprehensif dan pemrosesan teks, studi ini berhasil mengidentifikasi topik-topik kunci dalam narasi media terkait pemilu. Topik-topik tersebut mencakup dinamika pemilihan umum, seperti strategi kampanye dan pengaruh figur politik signifikan, serta elemen logistik dan administratif dari pemilu. Metode LDA, yang diperkaya dengan visualisasi interaktif menggunakan pyLDAvis, memungkinkan analisis detil tentang pendekatan media dalam membahas isu-isu ini. Hasil dari penelitian ini memberikan wawasan yang mendalam mengenai bagaimana berbagai aspek pemilu dipersepsikan dan dipresentasikan kepada publik melalui media. Temuan ini sangat relevan tidak hanya untuk memahami konteks pemilu secara lebih luas, tetapi juga dalam konteks merancang strategi komunikasi politik yang efektif untuk masa depan. Penelitian ini juga menyoroti pentingnya pemahaman tentang narasi media dalam menggambarkan gambaran politik, serta memfasilitasi diskusi yang lebih berinformasi mengenai kebijakan publik dan partisipasi politik di Indonesia.
Kata Kunci - Latent Dirichlet Allocation (LDA), Pemilu 2019, Pemodelan Topik.

I. Pendahuluan

Pemilu 2019 di Indonesia adalah salah satu peristiwa politik terbesar dalam sejarah negara ini. Pemilihan umum ini digelar pada tanggal 17 April 2019 dan melibatkan pemilihan presiden, anggota Dewan Perwakilan Rakyat (DPR), anggota Dewan Perwakilan Daerah (DPD), serta anggota Dewan Perwakilan Rakyat Daerah (DPRD) di berbagai bgvdf tingkatan provinsi, kabupaten, dan kota di seluruh Indonesia (Zuhro, 2019).

1. Pemilu 2019 menandai momen penting dalam perkembangan demokrasi di Indonesia, dengan lebih dari 190 juta pemilih terdaftar . Pemilihan ini juga ditandai oleh persaingan ketat antara calon presiden petahana, Joko Widodo (Jokowi), dan calon lawannya, Prabowo Subianto. Selama kampanye pemilu, isu-isu utama seperti ekonomi, infrastruktur, hak asasi manusia, dan kedaulatan ekonomi menjadi sorotan utama.

2. Tujuan penelitian ini adalah untuk mengkaji berita-berita tentang Pemilu 2019 yang tersebar dalam judul berita online. peneliti akan menggunakan metode Latent Dirichlet Allocation (LDA) untuk menganalisis dan mengklasifikasikan berita-berita ini ke dalam berbagai topik yang mungkin muncul dalam pemberitaan seputar pemilu tersebut.

3. Latent Dirichlet Allocation (LDA) adalah sebuah metode dalam analisis teks yang digunakan untuk mengidentifikasi topik-topik tersembunyi dalam koleksi dokumen . Metode ini dikembangkan oleh David Blei, Andrew Ng, dan Michael Jordan pada tahun 2003 . LDA bekerja dengan mengasumsikan bahwa setiap dokumen adalah gabungan dari beberapa topik, dan setiap kata dalam dokumen berasal dari salah satu topik tersebut .

4. Pendekatan Latent Dirichlet Allocation (LDA) dalam penelitian ini efektif dalam mengidentifikasi kata-kata yang sering muncul bersama dan mengelompokkannya ke dalam topik yang berbeda. Untuk menentukan jumlah topik yang paling tepat, penelitian ini menggunakan model coherence-gensim, dengan batasan perhitungan pada 21 model topik. Nilai coherence tertinggi ditentukan setelah melakukan serangkaian percobaan pemodelan awal.

5. Setelah dataset melewati tahap preprocessing dan cleaning data, langkah selanjutnya adalah pembentukan matriks dari kumpulan korpus, yaitu Bag of Word (BoW). Pemilihan fitur BoW ini penting untuk membuat pemodelan LDA menjadi lebih efektif dan relevan dengan tujuan penelitian. Hasil evaluasi topic coherence kemudian diproses menggunakan model LDA.

6. Proses ini tidak hanya mengidentifikasi topik-topik utama dalam dataset, tetapi juga memungkinkan visualisasi topik terbaik menggunakan modul pyLDAvis . Visualisasi ini memberikan wawasan tambahan tentang distribusi dan hubungan antar topik, serta kata-kata kunci yang paling mendefinisikan setiap topik. Dengan demikian, penelitian ini berhasil menerapkan metode LDA dan coherence-gensim untuk menghasilkan pemahaman yang lebih mendalam tentang struktur topik dalam dataset yang diteliti.

7. Beberapa penelitian tentang pemodelan topik pernah dilakukan sebelumnya. diantaranya penelitian yang dilakukan oleh yang menggunakan metode Latent Dirichlet Allocation (LDA) untuk menganalisis percakapan di media sosial tentang vaksinasi COVID-19 di Indonesia. Mereka menganalisis 1797 tweet, mengidentifikasi topik utama yang dibicarakan oleh pengguna Twitter. Hasil analisis menunjukkan beberapa topik penting seperti hak individu dalam memilih vaksinasi, kontroversi seputar tokoh publik Ribka Tjiptaning, dan penolakan vaksin oleh beberapa kelompok masyarakat. Penelitian ini memberikan pandangan yang mendalam tentang sentiment publik terhadap program vaksinasi COVID-19 di media sosial.

8. Selanjutnya, penelitian yang dilakukan oleh mengkaji aplikasi mobile "PeduliLindungi" yang dikembangkan oleh Pemerintah Indonesia sebagai langkah pencegahan pandemi COVID-19. Studi ini melakukan analisis pemodelan topik terhadap ulasan aplikasi di Google Play Store menggunakan metode Latent Dirichlet Allocation (LDA). Hasilnya adalah skor koherensi 0.3963 untuk total lima topik. Penelitian ini memberikan wawasan tentang persepsi dan reaksi masyarakat terhadap aplikasi "PeduliLindungi", yang mencerminkan pandangan publik terhadap inisiatif pemerintah dalam menangani pandemi.

9. Terakhir, penelitian yang dilakukan oleh berfokus pada portal berita online Detikcom. Penelitian ini menggunakan metode Latent Dirichlet Allocation (LDA) untuk melakukan ekstraksi topik dari berbagai berita yang terkumpul. Hasilnya adalah identifikasi tiga topik utama dengan skor koherensi 0,7586.

10. Berdasarkan latar belakang dan beberapa penelitian terdahulu yang telah disebutkan, penelitian ini bertujuan untuk menganalisis dan mengidentifikasi topik-topik utama yang muncul dalam pemberitaan terkait Pemilu 2019 di Indonesia. Penelitian ini bertujuan untuk memahami dinamika dan isu-isu yang paling banyak dibahas oleh media selama periode pemilihan, serta untuk melihat bagaimana berbagai narasi dan perspektif disajikan kepada publik.

11. 12. 13. 14. 15. 16. 17. 18. 19. 20. II. Metode

21. Beberapa langkah penting diadopsi dalam penelitian ini, menggunakan metode analisis Latent Dirichlet Allocation (LDA). Langkah-langkah tersebut meliputi:

22.

23.

24.

25.

26. Gambar 1 Alur Penelitian

27.

1. Pengambilan Dataset

Gambar 1 mengilustrasikan tahapan awal penelitian, yang melibatkan proses pengambilan dataset. Dataset ini diperoleh dari hasil scraping pada portal berita online. Informasi yang dihimpun dari portal tersebut kemudian disimpan dalam format Excel, yang memudahkan pengolahan dan analisis data selanjutnya.

2. Pra-pemorsesan Teks

Dataset ini pertama-tama harus dibersihkan melalui berbagai langkah, yang dimulai dengan tahap Pra-pemrosesan Teks yang ditunjukkan pada Gambar 1. Tahap ini meliputi pengubahan semua teks menjadi format huruf kecil. Proses tokenisasi bertujuan untuk memisahkan setiap kata dalam dokumen. Tahapan ini juga termasuk menghapus karakter tidak penting seperti emoji, tanda baca, link, tagar, dan URL. Setelah teks dibersihkan, langkah selanjutnya adalah menghilangkan kata-kata yang tidak memberikan makna atau pengaruh yang signifikan, dikenal sebagai penghapusan stopwords. Penghapusan stopwords ini berperan penting dalam meningkatkan proporsi informasi yang relevan dalam teks yang tidak terstruktur, sehingga meningkatkan kepentingan statistik dari istilah-istilah yang mungkin penting. Selain itu, penting untuk menghilangkan kata-kata berulang dalam dataset untuk memperoleh hasil terbaik saat pemodelan (normalisasi) serta melakukan pengurangan kata-kata dengan imbuhan melalui proses stemming.

3. Pemorsesan Teks

Gambar 1 menunjukkan langkah-langkah setelah tahap pra-pemrosesan teks, di mana data teks yang sudah dibersihkan dikonversi menjadi matriks Bag of Words (BOW) sebelum diproses lebih lanjut untuk pemodelan topik dan penimbangan kata-kata. Dalam metode BOW, data dalam bentuk tabel (korpus) diubah menjadi representasi numerik yang dikenal sebagai dokumen istilah. Proses ini hanya memperhitungkan keunikan kata; artinya, kata

yang sama atau duplikat yang muncul di berbagai topik hanya akan dicatat satu kali dalam matriks BOW.

1. **Latent Dirichlet Allocation (LDA) Latent Dirichlet Allocation (LDA) merupakan** model generatif berbasis probabilitas yang memandang setiap topik sebagai kumpulan dari berbagai kata atau 'token'. Model ini menganggap setiap dokumen atau korpus sebagai kombinasi dari berbagai topik probabilitas yang tersembunyi atau 'latent'. Salah satu keunggulan utama dari LDA adalah kapasitasnya dalam mengklasifikasikan dan mengolah data dalam skala besar. Hal ini terutama berkat cara LDA memberikan bobot atau penimbangan terhadap berbagai topik yang diidentifikasi dalam proses analisisnya, memungkinkan pemahaman yang lebih mendalam dan terstruktur tentang struktur topik dalam kumpulan data yang besar . Algoritma Latent Dirichlet Allocation (LDA) bekerja dengan menetapkan sejumlah parameter awal untuk menjalankan proses analisisnya. Parameter-parameter ini termasuk jumlah dokumen dalam kumpulan data (M), jumlah topik (K) yang ingin diidentifikasi dalam analisis, jumlah iterasi (i) yang akan dilakukan selama proses pemodelan, jumlah kata per dokumen (N), serta koefisien LDA (α, β). Koefisien α dan β **ini sangat penting karena mereka memainkan peran kunci** dalam **mengatur** distribusi topik dan kata. α berkaitan dengan distribusi topik di dalam dokumen, sedangkan β berkaitan dengan distribusi kata dalam topik. Penyesuaian parameter ini memungkinkan LDA untuk secara efektif menemukan struktur topik yang tersembunyi dalam kumpulan data dan menghasilkan pemahaman yang lebih mendalam tentang keterkaitan antara dokumen dan topiknya.

Gambar 2 Alur Kerja LDA

Latent Dirichlet Allocation (LDA) awalnya memberikan label kepada setiap kata dalam dokumen dengan topik secara acak, berdasarkan distribusi acak. Selama proses iterasi, LDA mengambil sampel ulang setiap kata ("resample each word") untuk memperbaiki parameter-parameter yang akan menentukan sebaran topik dalam dokumen tersebut. Proses sampel ulang ini didasarkan pada seberapa umum suatu kata muncul dalam suatu topik, serta seberapa sering topik itu muncul dalam dokumen.

Gambar 2 mengilustrasikan bagaimana distribusi Dirichlet, yang dikendalikan oleh parameter α, mempengaruhi distribusi topik dalam suatu dokumen dalam konteks model Latent Dirichlet Allocation (LDA). Nilai α yang lebih tinggi mengindikasikan bahwa dokumen cenderung mencakup berbagai topik, mencerminkan variasi yang lebih luas dalam isi topiknya. Hal ini diwujudkan melalui distribusi multinomial (θ), dimana setiap sampel dari distribusi ini menghasilkan kombinasi topik (Z) dan kata-kata spesifik (W) yang berbeda. Dalam model LDA, setiap dokumen (M) bisa terdiri dari jumlah kata (N) yang berbeda. Di sisi lain, nilai α yang lebih rendah menandakan bahwa dokumen tersebut cenderung lebih terfokus pada beberapa topik saja, menghasilkan sebaran topik yang lebih homogen dan tidak terlalu bervariasi. Gambar tersebut menggambarkan bagaimana LDA mengatur kompleksitas distribusi topik dalam dokumen berdasarkan nilai α, memungkinkan model untuk menyesuaikan dengan kekayaan dan keragaman topik dalam data yang dianalisis.

Dalam model Latent Dirichlet Allocation (LDA), parameter β mengendalikan bagaimana kata-kata tersebar dalam setiap topik (φ). Ketika nilai β tinggi untuk suatu topik, ini menunjukkan bahwa kata-kata dalam topik tersebut dapat ditemukan di berbagai topik lainnya. Sebaliknya, jika nilai β rendah, maka kata-kata dalam topik tersebut cenderung eksklusif dan lebih khusus untuk topik tersebut. Dalam LDA, yang diamati adalah distribusi kata-kata (W) dalam dokumen.

2. Coherence Score

Skor koherensi topik atau "topik koherensi" merupakan suatu metrik yang digunakan untuk mengevaluasi kualitas pemodelan topik . Sebuah model yang efektif akan menghasilkan topik-topik dengan tingkat koherensi yang tinggi. Saat menguji model ini, prosesnya melibatkan iterasi sebanyak jumlah topik yang ada, dimulai dari topik pertama. Koherensi topik ini memanfaatkan statistik dan probabilitas yang berasal dari korpus referensi, yaitu kumpulan teks, dengan fokus pada konteks kata untuk mengevaluasi koherensi setiap topik.

Pendekatan ini telah diakui sebagai metode evaluasi intrinsik yang efektif untuk model pemodelan topik. Skor koherensi topik dihitung dengan mempertimbangkan kesamaan kata yang berpasangan, yang didasarkan pada kata-kata paling penting dalam setiap topik sesuai dengan persamaan topik yang relevan. Semakin tinggi skor koherensi, semakin baik kualitas interpretasi dan relevansi topik yang dihasilkan oleh model. Ini mengindikasikan bahwa topik-topik yang dihasilkan memiliki kualitas koherensi yang tinggi dan mudah dimengerti dalam konteks yang lebih luas.

III. Hasil dan Pembahasan

1. Pengambilan Dataset

Data yang digunakan untuk analisis diambil dari portal berita online selama periode tahun 2017 hingga tahun 2022. Sebagai contoh, dataset awal dapat ditemukan pada Tabel 3.1, yang berisi data mentah yang belum mengalami proses pembersihan atau pengolahan lebih lanjut.

Tabel 3. SEQ Tabel * ARABIC 1 Sampel dataset

Teks
Kampanye Pilkada Bangkalan Mulai Tidak Sehat
Benarkah Polsek Tanjung Bumi Tak Gubris Laporan Tim Farid Alfauzi?
Bila Pendaftaran Dibuka, Farid Alfauzi Sudah Bisa Mendaftar Pilkada Bangkalan
Ratusan Baleho Bakal Calon Bupati Bangkalan Dirusak
Begini Cara Nyaleg Lewat Hanura Bangkalan

2. Pra-Pemrosesan Teks

Dataset yang diperoleh dari portal berita online tidak dapat langsung digunakan oleh sistem. Oleh karena itu, diperlukan beberapa tahap preprocessing atau pra-pemrosesan data untuk mengubahnya agar memenuhi standar dan meningkatkan kualitas data yang akan digunakan dalam analisis.

1. Lowercase Folding

Data awal yang merupakan teks mentah belum menjalani proses apapun, sehingga masih memiliki bentuk dan struktur asli. Dalam pemrosesan data teks, langkah awal yang umum adalah mengubah semua huruf dalam dokumen atau data menjadi huruf kecil. Ini dapat dilakukan dengan menggunakan fungsi str.lower() dalam Python . Fungsi ini tidak hanya mengkonversi teks ke huruf kecil, tetapi juga membantu menghilangkan karakter yang bukan huruf dari 'a' sampai 'z', termasuk delimiter dan karakter khusus lainnya. Hasil dari proses ini, yang dikenal sebagai "lowercase folding", dapat dilihat pada Tabel 3.2.1, di mana teks asli telah diubah menjadi format yang lebih seragam dan mudah untuk diolah lebih lanjut dalam analisis teks atau pemrosesan NLP.

Tabel 3.2.1. Perbandingan sebelum dan setelah casefolding

Sebelum casefolding	Setelah Casefolding
Kampanye Pilkada Bangkalan Mulai Tidak Sehat	kampanye pilkada bangkalan mulai tidak sehat
Benarkah Polsek Tanjung Bumi Tak Gubris Laporan Tim Farid Alfauzi?	benarkah polsek tanjung bumi tak gubris laporan tim farid alfauzi?
Bila Pendaftaran Dibuka, Farid Alfauzi Sudah Bisa Mendaftar Pilkada Bangkalan	bila pendaftaran dibuka, farid alfauzi sudah bisa mendaftar pilkada bangkalan

Ratusan Baleho Bakal Calon Bupati Bangkalan Dirusak ratusan baleho bakal calon bupati bangkalan dirusak
Begini Cara Nyaleg Lewat Hanura Bangkalan begini cara nyaleg lewat hanura bangkalan

1. Tokenizing

Setelah proses case folding, data akan dipecah menjadi segmen-segmen karakter seperti kata- **kata, tanda baca, angka, dan simbol.** Segmen-segmen ini dikenal **sebagai token.** Untuk memecah **kalimat dalam dokumen menjadi kata per** kata. Anda **dapat menggunakan** fungsi `word_tokenize()` **dengan mengimpor modul `nltk.tokenize` dan `nltk.probability` terlebih dahulu.** Selain itu, dalam tahap ini juga akan dilakukan penghapusan **tanda baca** menggunakan fungsi `remove_punctuation()`, **sehingga data** yang tersisa hanya **berupa** fragmen kalimat yang terdiri dari kata-kata saja.

Tabel SEQ Tabel * ARABIC 3.2.2. Perbandingan sebelum dan setelah tokenizing

Sebelum Tokenizing	Setelah Tokenizing
kampanye pilkada bangkalan mulai tidak sehat	['kampanye', 'pilkada', 'bangkalan', 'mulai', 'tidak', 'sehat']
benarkah polsek tanjung bumi tak gubris laporan tim farid alfauzi	['benarkah', 'polsek', 'tanjung', 'bumi', 'tak', 'gubris', 'laporan', 'tim', 'farid', 'alfauzi']
bila pendaftaran dibuka farid alfauzi sudah bisa mendaftar pilkada bangkalan	['bila', 'pendaftaran', 'dibuka', 'farid', 'alfauzi', 'sudah', 'bisa', 'mendaftar', 'pilkada', 'bangkalan']
ratusan baleho bakal calon bupati bangkalan dirusak	['ratusan', 'baleho', 'bakal', 'calon', 'bupati', 'bangkalan', 'dirusak']
begini cara nyaleg lewat hanura bangkalan	['begini', 'cara', 'nyaleg', 'lewat', 'hanura', 'bangkalan']

2. Stopwords

Dalam text mining, struktur kalimat yang kompleks seringkali **dianggap kurang** efektif untuk **pemodelan topik** dan analisis sentimen **karena dapat** mengurangi akurasi hasil **analisis. Oleh karena itu,** langkah penting yang dilakukan adalah menghilangkan kata-kata yang tidak bermakna atau stopwords yang sering muncul namun tidak memberikan informasi signifikan. Anda dapat melakukan ini dengan menggunakan modul `nltk.corpus`. Sebagai contoh, dalam bahasa Indonesia, beberapa stopwords yang sering dihapus menggunakan modul Sastrawi mencakup kata-kata seperti **" dan," "pada," "pd," "begitu," "gini," "gitu," "yang," "adalah," "jika," "maka," dan** lainnya. **Perbedaan antara data asli dan data yang telah** diproses untuk menghilangkan stopwords **ini dapat dilihat pada** Tabel 3.2.3.

Tabel 3.2.3. Perbandingan sebelum dan setelah stopwords **Sebelum** Stopwords Setelah Stopwords

['kampanye', 'pilkada', 'bangkalan', 'mulai', 'tidak', 'sehat']	['kampanye', 'pilkada', 'bangkalan', 'sehat']
['benarkah', 'polsek', 'tanjung', 'bumi', 'tak', 'gubris', 'laporan', 'tim', 'farid', 'alfauzi']	['polsek', 'tanjung', 'bumi', 'gubris', 'laporan', 'tim', 'farid', 'alfauzi']
['bila', 'pendaftaran', 'dibuka', 'farid', 'alfauzi', 'sudah', 'bisa', 'mendaftar', 'pilkada', 'bangkalan']	['pendaftaran', 'dibuka', 'farid', 'alfauzi', 'mendaftar', 'pilkada', 'bangkalan']
['ratusan', 'baleho', 'bakal', 'calon', 'bupati', 'bangkalan', 'dirusak']	['ratusan', 'baleho', 'calon', 'bupati', 'bangkalan', 'dirusak']
['begini', 'cara', 'nyaleg', 'lewat', 'hanura', 'bangkalan']	['nyaleg', 'hanura', 'bangkalan']

Gambar SEQ Gambar * ARABIC 3 Wordcloud

Hasil dari proses ini dapat divisualisasikan dengan menggunakan wordcloud. Dalam **visualisasi** ini, **kata-kata** akan ditampilkan **dalam** berbagai ukuran, di mana ukuran kata tersebut merepresentasikan frekuensi kemunculannya. Sebagai contoh, dalam hasil visualisasi yang ditunjukkan pada Gambar 3, kata "menolak" dan "vaksin" muncul dengan ukuran paling besar, menandakan bahwa kedua kata tersebut sering muncul dalam dokumen atau dataset yang dianalisis.

3. Text Normalization

Sebelum memulai normalisasi teks, langkah awalnya adalah mengidentifikasi data dari 'tweet_token' yang telah disimpan selama proses tokenisasi. Data ini identik dengan 'tweet_tokens_WSW' (Word Stop Words), memudahkan dalam memanggil fungsi saat melakukan normalisasi dan proses pengunduhan. Setelah berhasil diunduh, hasilnya akan dibaca ulang oleh mesin dan diinisialisasi. Tujuan utama normalisasi teks adalah mengurangi pengulangan karakter dalam dataset, sehingga menghindari duplikasi kata, sebagaimana yang terlihat dalam Tabel 3.2.4.

Tabel 3.2.4. Perbandingan sebelum dan setelah normalisasi teks

Sebelum normalisasi	Setelah normalisasi
['kampanye', 'pilkada', 'bangkalan', 'sehat']	['kampanye', 'pilkada', 'bangkalan', 'sehat']
['polsek', 'tanjung', 'bumi', 'gubris', 'laporan', 'tim', 'farid', 'alfauzi']	['polsek', 'tanjung', 'bumi', 'gubris', 'laporan', 'tim', 'farid', 'alfauzi']
['pendaftaran', 'dibuka', 'farid', 'alfauzi', 'mendaftar', 'pilkada', 'bangkalan']	['pendaftaran', 'dibuka', 'farid', 'alfauzi', 'mendaftar', 'pilkada', 'bangkalan']
['ratusan', 'baleho', 'calon', 'bupati', 'bangkalan', 'dirusak']	['ratusan', 'baleho', 'calon', 'bupati', 'bangkalan', 'dirusak']
['nyaleg', 'hanura', 'bangkalan']	['nyaleg', 'hanura', 'bangkalan']

4. Stemming

Dalam tahap ini, data yang telah melalui normalisasi, yang direpresentasikan oleh pemanggilan fungsi 'tweet_normalized', akan mengalami proses tambahan. Modul Python Sastrawi memainkan peran penting dalam tahap ini dengan menghapus imbuhan dan infleksi kata dalam bahasa Indonesia sehingga kata-kata tersebut berubah **ke** bentuk dasarnya, termasuk penghapusan **prefiks** dan **sufiks.** Tahap **akhir** dari proses **stemming** ini melibatkan penyamakan ' tweet_normalized' agar sesuai dengan fungsi 'tweet_token_stemmed'. Untuk memahami perbedaan antara data sebelum dan setelah proses stemming, Anda dapat merujuk ke Tabel 3.2.5.

Tabel 3.2.5. Perbandingan sebelum dan setelah stemming

Sebelum Stemming	Setelah Stemming
['kampanye', 'pilkada', 'bangkalan', 'sehat']	['kampanye', 'pilkada', 'bangkal', 'sehat']
['polsek', 'tanjung', 'bumi', 'gubris', 'laporan', 'tim', 'farid', 'alfauzi']	['polsek', 'tanjung', 'bumi', 'gubris', 'lapor', 'tim', 'farid', 'alfauzi']
['pendaftaran', 'dibuka', 'farid', 'alfauzi', 'mendaftar', 'pilkada', 'bangkalan']	['daftar', 'buka', 'farid', 'alfauzi', 'daftar', 'pilkada', 'bangkal']
['ratusan', 'baleho', 'calon', 'bupati', 'bangkalan', 'dirusak']	['ratus', 'baleho', 'calon', 'bupati', 'bangkal', 'rusak']
['nyaleg', 'hanura', 'bangkalan']	['nyaleg', 'hanura', 'bangkal']

5. Pembuatan Bag of Word (BOW)

Setelah menyelesaikan proses pra-pemrosesan teks, langkah selanjutnya adalah mengubah dokumen dari dataset menjadi vektor yang dikelompokkan

dalam satu kelompok. Model Bag of Words (BoW) digunakan untuk mengklasifikasikan teks dan mengekstrak berbagai fitur dari teks tersebut (Qiu et al., 2020). Pendekatan ini beroperasi dengan menghitung frekuensi kemunculan setiap kata dalam satu dokumen. Dengan cara ini, diperoleh Bag of Words untuk dokumen dengan indeks ke-394, yang dapat dilihat dengan lebih detail dalam Tabel 3.2.6.

Tabel 3.2.6. Frekuensi **Kemunculan Kata Kata Frequency Kemunculan**

milu 1
warga 1
kpu 1
ajak 1
nyoblos 1
run 1

1. **Latent Dirichlet Allocation (LDA)**

Model Latent Dirichlet Allocation (LDA) memungkinkan penentuan **topik yang diringkas secara** subjektif. Namun, **untuk** menginterpretasikan topik-topik **yang diperoleh secara** maksimal, dilakukan **perhitungan koherensi** sebelum memulai **pemodelan topik.** Proses perhitungan koherensi **ini** memerlukan sekumpulan **korpus yang telah** disiapkan sebelumnya, serta kamus kata-kata yang sesuai. Langkah ini sangat penting untuk memastikan bahwa topik yang dihasilkan oleh model LDA memiliki tingkat relevansi dan konsistensi yang tinggi, sehingga meningkatkan kualitas dan akurasi interpretasi dari topik yang diidentifikasi.

Gambar 4 Hubungan Topik dengan Coherence Score

Grafik koherensi score yang dihasilkan dari model Latent Dirichlet Allocation (LDA) menunjukkan informasi penting mengenai distribusi topik dalam dataset. Skor koherensi yang meningkat secara signifikan hingga puncak pada 6 topik menandakan bahwa pada jumlah topik ini, model LDA berhasil mengidentifikasi topik dengan kejelasan dan distingsi yang lebih baik dibandingkan dengan jumlah topik yang lebih rendah. Ini mengindikasikan bahwa setiap topik memiliki kata kunci yang cukup spesifik dan relevan, sehingga memudahkan interpretasi dan pengelompokan dokumen.

Setelah titik puncak pada 6 topik, meskipun terjadi peningkatan koherensi secara bertahap hingga 20 topik, kenaikan tersebut tidak lagi signifikan dan cenderung stabil. Hal ini dapat mengimplikasikan bahwa penambahan topik di atas enam tidak secara substansial meningkatkan kemampuan model untuk menangkap struktur semantik yang lebih kaya dalam data. Oleh karena itu, dalam konteks skripsi ini, pemilihan 6 topik tampaknya menjadi pilihan optimal, memberikan keseimbangan antara kompleksitas model dan kejelasan tematik yang diperoleh.

Dalam pengembangan lebih lanjut dari penelitian ini, akan diambil keputusan untuk memodelkan topik menggunakan 6 topik yang telah ditentukan sebagai jumlah optimal. Keputusan ini didasarkan pada bukti yang diperoleh dari grafik koherensi yang menunjukkan bahwa 6 topik menghasilkan koherensi tertinggi, yang berarti interpretasi topik akan lebih koheren dan akurat. Ini juga memastikan bahwa model tidak overfit atau underfit, mempertahankan generalisasi yang sesuai tanpa mengorbankan detil spesifik dari data.

Gambar 5 Visualisasi PyLDAvis

Dalam visualisasi pyLDAvis yang disediakan, Intertopic Distance Map menunjukkan posisi dan hubungan antar topik dalam model Latent Dirichlet Allocation (LDA). Topik nomor 13, yang ditandai dengan warna merah, terletak agak terpisah dari topik lainnya, menandakan bahwa topik ini memiliki ciri khas yang lebih unik dan kurang tumpang tindih dengan topik lain dalam model. Hal ini bisa menunjukkan adanya set kata-kata yang sangat khusus atau diskusi yang sangat berfokus dalam korpus yang diteliti.

Di sisi lain, daftar 'Top-30 Most Relevant Terms for Topic 13' mengungkapkan **kata-kata yang paling dominan dalam topik** tersebut, dengan "jokowi", "unggul", dan "malang" menjadi yang paling menonjol. Frekuensi relatif dari istilah-istilah ini, yang ditunjukkan dengan panjang bar biru, memberikan indikasi tentang istilah mana yang paling berpengaruh dalam membentuk konten topik tersebut. Kehadiran istilah-istilah seperti "politik", "prabowo", "survey", dan "korupsi" bisa mengindikasikan bahwa topik ini terkait erat dengan pembahasan politik di Indonesia.

Distribusi marginal topik menunjukkan bahwa Topik 13 mencakup 4.9% dari total token dalam model, yang menandakan bahwa ini adalah topik yang cukup signifikan dalam korpus. Dalam konteks skripsi, analisis ini dapat memberikan pemahaman yang lebih mendalam mengenai aspek-aspek tertentu dari diskusi politik, yang dapat menjadi fokus analisis lebih lanjut atau digunakan untuk memperkuat argumentasi dalam penelitian.

INCLUDEPICTURE "https://files.oaiusercontent.com/file-zUzloZ3T8hERXIXckEdbXfy5?se=2024-01-30T02%3A32%3A09Z&sp=r&sv=2021-08-06&sr=b&rscc=max-age%3D299%2C%20immutable&rscd=attachment%3B%20filename%3Dimage.png&sig=fB5V9u4fBOrgqZVfo100Q8Y8fTw/8fX2rHO2uwrdrOA%3D" * MERGEFORMATINET

Gambar 6 Wordcloud Setiap Topik

Word cloud yang dihasilkan dari pemodelan LDA mengungkapkan kosa kata kunci yang membentuk lima topik berbeda dalam korpus data yang dianalisis. Topik 0, yang diwarnai hijau, menyoroti prosedur pemilu dengan istilah seperti "nyoblos" dan "kpu", menandakan fokus pada aktivitas pemilihan umum, dengan referensi spesifik ke lokasi seperti Malang dan Pacitan. Topik 1, berwarna merah, berkonsentrasi pada pemilihan kepala daerah, dengan istilah "pilkada" dan "calon", serta "rawan" yang menyarankan adanya risiko atau tantangan dalam konteks pemilihan di Jawa Timur. Topik 2, yang ditampilkan dalam warna biru tua, mencakup terminologi yang berkaitan dengan partai politik seperti "penuh", "parpol", "golkar", menunjukkan diskusi tentang struktur dan kebijakan partai, serta "sandiga" yang mungkin mengacu pada tokoh politik Sandiaga Uno. Topik ini menggambarkan aspek-aspek kepartaian dan kriteria keanggotaan atau keterlibatan dalam politik domestik.

Topik 3, berwarna oranye, memuat nama-nama penting dalam politik seperti "jokowi" dan "prabowo", serta kata-kata seperti "politik" dan "survey", menandakan analisis tentang pengaruh dan dukungan politik mereka. Topik 4, dengan warna ungu, merinci aspek administratif dan logistik pemilu dengan kata-kata seperti "pkpu" dan "dpt", menunjukkan fokus pada regulasi dan manajemen pemilihan.

Akhirnya, Topik 5, dalam warna biru muda, sepertinya mengarah pada prosedur pemungutan dan penghitungan suara, dengan istilah-istilah seperti "rayalelang" dan "kardus" yang bisa jadi merujuk pada proses pemilihan itu sendiri. Setiap topik menawarkan pandangan unik tentang aspek yang berbeda dari dialog politik, mulai dari mekanisme pemilu hingga diskusi tentang figur politik dan kebijakan partai.

IV. Simpulan

Setelah menganalisis data dan mengolahnya melalui berbagai tahap pemrosesan teks dan pemodelan LDA, penelitian ini telah berhasil mengidentifikasi topik-topik utama yang muncul dalam pemberitaan terkait Pemilu 2019 di Indonesia. Topik-topik yang teridentifikasi mencakup dinamika pemilihan umum, fokus pada figur politik spesifik, serta aspek-aspek logistik dan administratif yang berkaitan dengan pemilihan. Dengan menggunakan model LDA dan

alat visualisasi seperti pyLDAvis, penelitian ini memberikan wawasan mendalam tentang cara topik-topik tersebut ditangani dan dibahas dalam media. Hasil ini menawarkan pandangan yang berharga tentang bagaimana berbagai aspek dari pemilu dipersepsikan dan disajikan kepada publik, yang tidak hanya penting untuk memahami konteks pemilu tersebut tetapi juga untuk merancang strategi komunikasi politik yang lebih efektif di masa depan.

Ucapan Terima Kasih

Semoga karya tulis ini memberrikan manfaat dan pemahaman yang mudah dimengerti bagi para pembaca. Terima kasih kepada semua pihak yang telah berperan serta dalam proses penulisan dan penyusunan tulisan ini. Dukungan serta doa yang diberikan oleh berbagai pihak selama tahap penelitian dan penulisan sangat berarti.

Referensi

[1] Akbar, J., M., T. A., Tolla, Y., Ahmad, A., E., Yaqin, A., & Utami, E. (2023). Pemodelan Topik Menggunakan Latent Dirichlet Allocation pada Ulasan Aplikasi PeduliLindungi. InComTech: Jurnal Telekomunikasi Dan Komputer, 13(1), 40. https://doi.org/10.22441/incomtech.v13i1.15572

[2] Febrianta, M. Y., Widiyanesti, S., & Ramadhan, S. R. (2021). Analisis Ulasan Indie Video Game Lokal pada Steam Menggunakan Analisis Sentimen dan Pemodelan Topik Berbasis Latent Dirichlet Allocation. Journal of Animation and Games Studies, 7(2). https://doi.org/10.24821/jags.v7i2.5162

[3] Fernanda, J. W. (2021). PEMODELAN PERSEPSI PEMBELAJARAN ONLINE MENGGUNAKAN LATENT DIRICHLET ALLOCATION. Jurnal Statistika Universitas Muhammadiyah Semarang, 9(2). https://doi.org/10.26714/jsunimus.9.2.2021.79-85

[4] Istianda, M., & Zastrawati, A. (2021). EVALUASI PENYELENGGARAAN PEMILU SERENTAK 2019 KOTA MAKASSAR. Sebatik, 25(1). https://doi.org/10.46984/sebatik.v25i1.1203

[5] Matira, Y., & Setiawan, I. (2023). Pemodelan Topik pada Judul Berita Online Detikcom Menggunakan Latent Dirichlet Allocation. Estimasi: Journal of Statistics and Its Application, 4(1), 2721-379. https://doi.org/10.20956/ejsa.vi.24843

[6] Mifrah, S. (2020). Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus. International Journal of Advanced Trends in Computer Science and Engineering, 9(4). https://doi.org/10.30534/ijatcse/2020/231942020

[7] Mohammed, S. H., & Al-Augby, S. (2020). LSA & LDA topic modeling classification: Comparison study on E-books. Indonesian Journal of Electrical Engineering and Computer Science, 19(1). https://doi.org/10.11591/ijeecs.v19.i1.pp353-362

[8] Qiu, D., Jiang, H., & Chen, S. (2020). Fuzzy information retrieval based on continuous bag-of-words model. Symmetry, 12(2). https://doi.org/10.3390/sym12020225

[9] Sholihah, U. M., Findawati, Y., & Indahyanti, U. (2023). TOPIC MODELING IN COVID-19 VACCINATION REFUSAL CASES USING LATENT DIRICHLET ALLOCATION AND LATENT SEMANTIC ANALYSIS. 4(5), 1063-1074. https://doi.org/10.52436/1.jutif.2023.4.5.951

[10] Yang, H., Li, J., & Chen, S. (2023). TopicRefiner: Coherence-Guided Steerable LDA for Visual Topic Enhancement. IEEE Transactions on Visualization and Computer Graphics. https://doi.org/10.1109/TVCG.2023.3266890

[11] Zuhro, R. S. (2019). Demokrasi dan Pemilu Presiden 2019. Jurnal Penelitian Politik, 16(1). https://doi.org/10.14203/jpp.v16i1.782