FORMAT: Jurnal Ilmiah Teknik Informatika
Akreditasi Sinta 5 RISTEKBRIN No. 164/E/KPT/2021; E-ISSN 2722-7162 | P-ISSN 2089-5615

Vol. 13 , No. 1,
January 2024,
44 - 54. 2024

Komparasi Algoritma Topic Modelling LDA VS LSA Pada Berita Detikcom

Ahmad Kemal Al Izzi [1]; Rakadian Audiga Pratama [2]

[12] *Fakultas Sains dan Teknologi, Universitas Islam Negeri Sunan Ampel Surabaya, Jl. Dr. Ir. H. Soekarno No.682 Gn. Anyar Kec. Gn. Anyar, Surabaya, Jawa Timur 60294*

[1] kemalizzi213@gmail.com, [2]rakadianp@gmail.com,

Kata kunci:
Buah Pepaya,Fitur Warna,Klasifikasi,Metode SVM,Kematangan

Abstract

This research focuses on the process of applying Topic Modeling by comparing the Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) models on news tweet data taken from the Detikcom account. The process begins by crawling data over a one year period, starting from December 9, 2022 to December 9, 2023, resulting in 958 rows of data. Data pre-processing includes steps such as case folding, tokenization, stopwords removal, and stemming. After pre-processing, a bag of words process is carried out to calculate the frequency of word occurrences in each document. The number of word occurrence frequencies is used as a reference in creating LSA and LDA models. Each model has 8 topics, 10 iterations, and 42 random states. Topic production is carried out based on keywords that appear in the modeling results. Evaluation of the two models is carried out by measuring topic coherence or topic coherence using the c_v value. The LSA model shows a coherence value of 0.5, while the LDA model has a coherence value of 0.45. The evaluation results show that in this case, the LSA model has better performance than the LDA model based on the topic coherence value. As a suggestion for further research, researchers are expected to consider the use of other cases for topic modeling and other exploration models in Topic Modeling such as OCTIS. This can expand understanding of the performance of the Topic Modeling algorithm on X news data.

Pendahuluan

A. Latar Belakang

Perkembangan teknologi informasi dan komunikasi yang pesat telah membawa dampak yang signifikan dalam menyediakan akses informasi secara cepat, luas, dan global melalui internet. Banyak orang memanfaatkan Internet, mulai dari sektor industri, teknologi, dan lain-lain [1]. Globalisasi terus membawa perubahan pada banyak bidang di kehidupan nyata, salah satunya adalah bidang informasi dan komunikasi. Salah satu sumber informasi yang sangat populer saat ini adalah portal berita *online*. Pengguna portal berita online dapat mengakses berita terkini dari berbagai bidang seperti politik, hiburan, ekonomi, dan lain-lain secara *real-time* dimanapun dan kapanpun. Detikcom, sebagai salah satu portal berita online terkemuka di Indonesia, menjadi sumber utama masyarakat untuk mendapatkan informasi terkini.

Dengan pertumbuhan dan meningkatnya jumlah berita di era digital ini, mengekstraksi informasi yang relevan dan memahami topik-topik berita yang dominan menjadi semakin penting. Dalam hal ini, algoritma Topic Modelling menjadi instrumen yang cocok untuk mengidentifikasi tematik yang terdapat dalam koleksi berita. Dua di antara algoritma Topic Modelling yang paling umum digunakan adalah Latent Dirichlet Allocation (LDA) dan Latent Semantic Analysis (LSA).

Penelitian ini bertujuan untuk melakukan komparasi antara algoritma Topic Modelling LDA dan LSA pada berita-berita yang dipublikasikan oleh Detikcom. Dengan menganalisis dan membandingkan hasil ekstraksi topik dari kedua algoritma, diharapkan dapat ditemukan keunggulan dan kelemahan masing-masing algoritma dalam konteks aplikasi berita online. Pentingnya penelitian ini terletak pada kontribusi terhadap pengembangan teknik *Text Analytics* dan pengelolaan informasi di era digital. Hasil dari penelitian ini dapat memberikan pandangan yang lebih jelas mengenai performa LDA dan LSA dalam mengidentifikasi topik-topik yang relevan pada akun portal berita Detikcom. Hal ini dapat menjadi landasan untuk peningkatan kualitas sistem rekomendasi berita, pengelompokan berita, dan pemahaman konten berita secara lebih mendalam.

Penelitian ini akan dibagi menjadi beberapa bagian, termasuk tinjauan pustaka terkait Topic Modelling, penjelasan detail mengenai algoritma LDA dan LSA, metodologi penelitian, hasil eksperimen, dan analisis hasil. Dengan merinci langkah-langkah penelitian secara sistematis, diharapkan dapat memberikan kontribusi yang signifikan dalam memahami perbandingan kinerja antara LDA dan LSA dalam konteks analisis berita online.

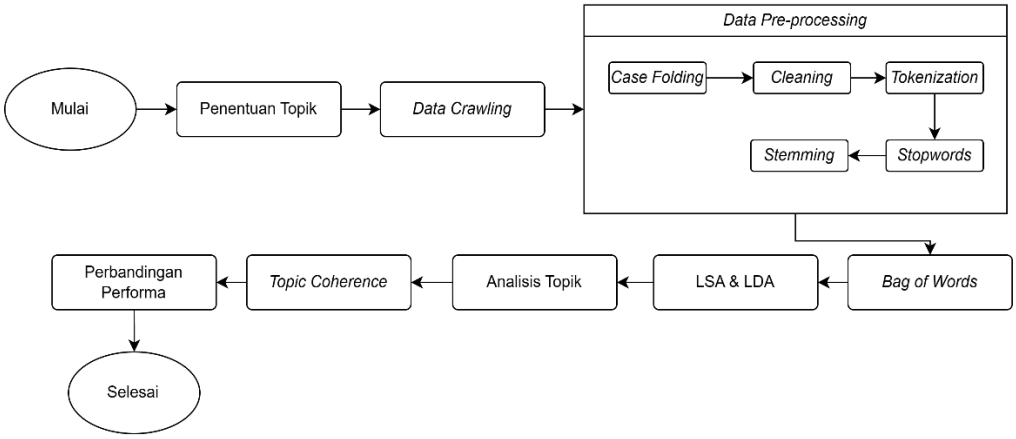
Penulis melaksanakan penelitian dengan tujuan mengembangkan model yang mampu menganalisis topik berdasarkan hasil analisis sentimen pada portal berita *online* Detikcom. Topik-topik yang dihasilkan akan dikelompokkan berdasarkan sentimen yang dihasilkan oleh mesin. Penelitian ini memberikan kontribusi dalam memperkaya literatur penelitian di bidang pemodelan topik. Selain itu, kontribusi lainnya adalah menyediakan wawasan yang berharga bagi masyarakat dan pengambil keputusan, memungkinkan mereka untuk memahami topik-topik yang dibahas dalam media massa terkait dengan suatu isu berita [2]. Diharapkan wawasan yang diperoleh dapat menjadi panduan bagi masyarakat dan pemangku kebijakan dalam proses pembuatan kebijakan terkait penanganan isu berita tersebut.

Metode penelitian

Penelitian kali ini akan menggunakan dua metode, yaitu *Latent Dirichlet Allocation* (LDA) dan *Latent Semantic Analysis* (LSA) yang nantinya performa dari kedua metode tersebut akan dibandingkan. Sedangkan metode perhitungan hasil penelitian nantinya akan menggunakan metode kuantitatif melalui *evaluation metrics*. *Evaluation metrics* adalah ukuran atau kriteria yang digunakan untuk mengevaluasi kinerja atau efektivitas suatu model atau sistem. Dalam konteks evaluasi model, metrics memberikan gambaran tentang seberapa baik model tersebut dapat memprediksi atau mengklasifikasikan data. Metrics ini digunakan untuk mengukur sejauh mana model mampu memberikan hasil yang akurat, konsisten, dan sesuai dengan tujuan yang diinginkan.

Dalam bidang *machine learning* atau *data mining*, *evaluation metrics* digunakan untuk mengukur performa model prediktif atau klasifikasi. Hasil prediksi merupakan output dari suatu program, dengan nilai yang dapat berupa Positif atau Negatif, sementara nilai aktual adalah nilai yang sebenarnya, dapat berupa *True* atau *False*[3]. Contoh beberapa *evaluation metrics* meliputi akurasi (*accuracy*), presisi (*precision*), *recall*, *F1-score*, area di bawah kurva ROC (AUC-ROC), dan sebagainya. Setiap metric memiliki kegunaan dan interpretasi tersendiri, tergantung pada jenis masalah dan tujuan evaluasi yang diinginka

Dalam evaluasi model, pemilihan *evaluation metrics* haruslah sesuai dengan konteks masalah dan kebutuhan pengguna. Hal ini membantu dalam memberikan pandangan yang holistik terhadap kinerja suatu model dan memastikan bahwa model tersebut dapat memberikan solusi yang efektif sesuai dengan kebutuhan yang dihadapi. Data yang digunakan berasal dari akun portal berita Detikcom yang ada di *platform X*. Alur penelitian ini dapat dilihat pada gambar:



Gambar 1. *Flowchart* penelitian.

A. Penentuan Topik

Topik yang akan dibahas pada penelitian kali ini adalah pemodelan topik pada akun portal berita Detikcom di *platform X*. Seperti pembahasan sebelumnya, dibutuhkan klasifikasi topik pada berita-berita yang diterbitkan oleh Detikcom agar pemahaman konten berita menjadi lebih dalam. Dengan tidak adanya pengklasfikasian berita di *platform X*, penelitian ini bertujuan untuk mengidentifikasi dan menganalisis hal tersebut. Pemahaman terhadap objek penelitian dilakukan dengan mencari dan menggali informasi dari sumber objek, sehingga proses ini bertujuan untuk menghasilkan hasil sesuai dengan harapan yang telah ditetapkan [4].

B. Data Crawling

Pengumpulan atau ekstraksi data dilakukan melalui metode *crawling*. *Data crawling* memiliki peran yang berbeda dengan *data scraping* kendati sama -sama digunakan pada halaman web. *Crawling* merupakan suatu metode untuk mengakses dan mengambil konten informasi yang terdapat pada sebuah situs web [5]. Data crawling menjadi metode yang sangat penting untuk mengumpulkan data dengan cepat, melihat pesatnya pertumbuhan pengguna dan konten di *WWW (World Wide Web)* karena mampu melakukan pencarian dan pengambilan data secara efisien [6]. Teknik pengumpulan informasi dapat mencakup pengambilan data secara menyeluruh atau dengan fokus khusus pada data tertentu.

C. Data Pre-processing

Pada langkah *pre-processing*, dokumen diubah menjadi format yang memudahkan dan mempercepat proses dalam mengidentifikasi dokumen yang relevan. Setiap tahap *pre-processing* bertujuan untuk membangun indeks dari kumpulan dokumen. Pengindeksan dilakukan untuk membedakan satu dokumen dari yang lain. Hal ini bertujuan agar hasil klasifikasi yang diinginkan dapat mencapai performa maksimal dan mengurangi risiko kesalahan [7]. Pembuatan indeks melibatkan konsep pemrosesan linguistik yang bertujuan untuk mengekstrak istilah-istilah penting dari setiap dokumen yang direpresentasikan sebagai "*bag of words*" (kumpulan kata). Konsep pemrosesan linguistik mencakup tokenisasi, penghilangan kata-kata penghenti (*stopwords*), dan stemming. Proses tersebut merupakan empat langkah dalam *pre-processing*. adapun rincian dari tiap proses pada *preprocessing* sebagai berikut:

1) Case Folding

Case folding adalah langkah yang melibatkan transformasi seluruh huruf dalam suatu dokumen atau kalimat menjadi huruf kecil. Tujuannya adalah untuk menyederhanakan proses pencarian karena konsistensi dalam penggunaan huruf kapital tidak selalu dijamin dalam data [8].

2) Cleaning

Cleansing adalah tahapan yang bertujuan mengeliminasi atribut-atribut yang tidak relevan, contohnya seperti URL, *mention, username*, RT, *hashtag*, atau tanda baca, dan menentukan kelas. Langkah ini akan mengubah teks menjadi data yang lebih mudah diproses oleh sistem selama tahap utama [9].

3) Tokenization

Tokenization atau Tokenisasi adalah tahap pemisahan suatu dokumen menjadi bagian-bagian yang disebut sebagai token [10].

4) Stopwords

Stopword adalah kata-kata yang tidak bernilai yang perlu dihilangkan. *Stopwords removal* adalah langkah untuk mengambil kata-kata kunci dari hasil token dengan menerapkan algoritma *stoplist* (menghilangkan kata-kata kurang penting) atau *wordlist* (menyimpan kata-kata kunci). Tujuan tahap ini untuk mengurangi dimensi dokumen teks input sehingga proses meringkas dapat dilakukan dengan lebih mudah [11].

5) Stemming

Stemming adalah proses di mana setiap kata diubah dari bentuk kata berimbuhan menjadi kata dasar [12]. Proses ini bertujuan untuk menyeragamkan kata sehingga dapat mengurangi daftar kata pada data *train*.

D. Bag of Words

Bag of Words adalah suatu model yang digunakan dalam *Natural Language Processing*, sering digunakan untuk ekstraksi nilai kata-kata yang telah diproses sebelumnya dalam suatu model *machine learning* [13]. Model *bag of words* menggambarkan setiap dokumen dengan tidak memperhatikan urutan kata-kata dan struktur sintaksis dari dokumen serta kalimat. Perhitungan dilakukan berdasarkan jumlah kemunculan setiap kata, yang kemudian digunakan dalam *topic modelling*.

E. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) merupakan suatu model probabilistik generatif yang digunakan untuk menganalisis kumpulan data teks. LDA dapat dianggap sebagai hierarki tiga tingkat dari model Naïve Bayes yang bertujuan untuk mengelompokkan kata-kata menjadi kluster, yang disebut sebagai topik, dalam suatu dokumen [14]. Pada penelitian ini, LDA digunakan sebagai metode pengolahan topik untuk menganalisis data yang telah diambil, membantu mengidentifikasi topik yang muncul dalam sejumlah ulasan yang telah dikumpulkan [15]. Pemilihan LDA dilakukan karena kemampuannya dalam menganalisis data dan dokumen berukuran besar. LDA menggunakan metode *bag of words* untuk mengidentifikasi informasi topik yang tersembunyi dalam kumpulan dokumen besar [16]. Pendekatan ini memperlakukan setiap dokumen sebagai vektor jumlah kata dan menyajikan distribusi probabilitas untuk beberapa topik, sementara setiap topik diwakili sebagai distribusi probabilitas untuk beberapa kata. Mekanisme operasi LDA terdiri dari dua tahap utama: inferensi, yang menentukan bobot setiap kata dalam setiap dokumen dalam korpus, dan implementasi, di mana aplikasi LDA memenuhi kebutuhan temu kembali informasi.

F. Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) adalah teori dan pendekatan yang digunakan untuk mengekstrak dan mewakili makna kata secara kontekstual melalui perhitungan statistik pada kumpulan teks. LSA memfokuskan perhatian pada kata-kata kunci dalam suatu kalimat tanpa mempertimbangkan karakteristik linguistiknya. Dalam LSA, kata-kata diwakili dalam sebuah matriks semantik dan kemudian diolah secara matematis dengan menggunakan teknik *Singular Value Decomposition (SVD)* dari aljabar linier. SVD merupakan metode untuk mengidentifikasi dan mengurutkan dimensi yang menunjukkan variasi data yang paling signifikan, memungkinkan pendekatan yang lebih efektif pada data asli dengan dimensi yang lebih rendah [17]. LSA adalah suatu metode statistika murni, bukan merupakan bentuk pemrosesan dari bahasa alami atau *artificial intelligence*. LSA tidak mengandalkan susunan *semantic-network*, kamus, ataupun *knowledge-based*. Metode ini hanya memanfaatkan input teks biasa yang diuraikan menjadi kata-kata, kemudian didefinisikan sebagai string karakter unik, dan dipisahkan menjadi bagian bermakna seperti kalimat atau paragraph [18].

G. Analisis Topik

Pada tahap analisis topik, dilakukan evaluasi berdasarkan data output dari tahap sebelumnya. *Output* tersebut berupa grafik yang mencerminkan hasil penelitian dengan topik tertentu. Analisis topik dilakukan secara subjektif dengan menelaah data keluaran yang terdiri dari kumpulan kata yang membentuk topik. Selanjutnya, setiap dokumen disesuaikan dengan data keluaran yang mencakup dokumen dengan topik tersebut. Proses ini menghasilkan deskripsi topik yang informatif, memberikan gambaran representatif terkait isi dari masing-masing topik.

H. Topic Coherence

Pada kenyataannya, keberadaan topik sendiri tidak menjamin interpretasi yang jelas. Oleh karena itu, diperlukan pengukuran koherensi untuk membedakan antara topik. Oleh karena itu, dilakukan pengukuran koherensi untuk secara otomatis menilai dan menampilkan sejauh mana suatu topik dapat dianggap koheren. Suatu topik dianggap koheren jika hampir semua kata atau kata-kata dengan nilai tertinggi saling terkait. Untuk mencapai hasil koherensi ini, dilakukan pengukuran menggunakan konsep *topic coherence* dengan melakukan eksplorasi secara sistematis dan empiris dalam skala besar terhadap urutan kepentingan topik yang telah ditentukan oleh manusia [19]. Pengukuran *topic coherence* merupakan alat evaluasi yang dapat membantu memisahkan antara kata-kata yang memiliki interpretasi semantik dan statistik [20].

I. Perbandingan Performa

Bagian ini akan menjelaskan dan membandingkan bagaimana hasil performa kedua metode dalam topic modelling, yaitu LDA dan LSA. Perbandingan akan diukur menggunakan coherence score masing-masing metode. Hal itu karena coherence score mampu menunjukkan tingkat akurasi koherensi topik-topik yang diteliti. Semakin tinggi nilai coherence score pada suatu metode maka metode tersebut semakin baik.

Hasil dan diskusi

Bagian ini merupakan tahap di mana hasil pekerjaan dianalisis untuk memberikan solusi atau jawaban yang berkaitan dengan masalah yang diteliti. Dengan memanfaatkan data yang telah diperoleh, sampel penelitian diproses untuk melakukan analisis kesimpulan dan pembahasan.

A. Pengumpulan Data

Tahap awal dari penelitian ini dimulai dengan melakukan pengumpulan data. Pengumpulan data dilakukan dengan *crawling* pada media sosial X (dulu disebut Twitter). *Crawling data* ini menghasilkan data sebanyak 958 baris. Data diambil dari cuitan *tweet* berita dari akun @Detikcom dengan rentang waktu 1 tahun dimulai dari 9 Desember 2022 sampai 9 Desember 2023. Teknik *crawling data* ini menggunakan *tools* Tweet Harvest yang merupakan sebuah alat untuk *crawl data* dari media sosial X sehingga dapat memperoleh informasi dari *thread* atau *tweet* maupun komentar.

B. Preprocessing Data

Preprocessing data dilakukan dengan tujuan untuk mempersiapkan data sebelum nantinya dilakukan proses lebih lanjut. Tahap *prepocessing* atau prapremrosesan data ini dilakukan di Google Collabboratory. Tahap ini memerlukan beberapa pustaka (*library*) yang ada pada python seperti RegEx, NLTK, dan Sastrawi Tahapan dalam proses ini terpecah menjadi beberapa langkah, yaitu:

a. Case Folding

Tahap pertama dalam *preprocessing* dengan mengubah setiap kalimat menjadi huruf kecil atau disebut dengan *case folding.* Tujuan dari tahapan ini menyetarakan kata agar kata yang sama namun memiliki penulisan berbeda dalam penggunaan kapital tidak dianggap sebagai kata yang berbeda. Contoh beberapa hasil dari tahapan ini dijelaskan dalam tabel berikut :

Tabel 1. Hasil proses *case folding*

Sebelum *Case Folding*	Sesudah *Case Folding*
Chelsea masih belum konsisten dalam meraih hasil bagus di atas lapangan. Manajer Chelsea Mauricio Pochettino menyebut para pemainnya merasakan tekanan. https://t.co/TC32gucbWI	chelsea masih belum konsisten dalam meraih hasil bagus di atas lapangan. manajer chelsea mauricio pochettino menyebut para pemainnya merasakan tekanan. https://t.co/tc32gucbwi
Jude Bellingham tak sepenuhnya fit usai mengalami cedera bahu bulan lalu. Pelatih Real Madrid, Carlo Ancelotti, tak merisaukan kondisi pemainnya tersebut. https://t.co/13xM0vpqAM	jude bellingham tak sepenuhnya fit usai mengalami cedera bahu bulan lalu. pelatih real madrid, carlo ancelotti, tak merisaukan kondisi pemainnya tersebut. https://t.co/13xm0vpqam

b. Cleaning

Tahapan ini juga biasa disebut pembersihan data. Kata-kata yang tidak berarti, misalnya tanda baca, URL, *username*, *mention*, emoji, angka, dan simbol, dihapus pada langkah ini. Pada tahap ini memerlukan pustaka RegEx yang tersedia dalam python.

Tabel 2. Hasil proses *cleaning*

Sebelum *Cleaning*	Sesudah *Cleaning*
chelsea masih belum konsisten dalam meraih hasil bagus di atas lapangan. manajer chelsea mauricio pochettino menyebut para pemainnya merasakan tekanan. https://t.co/tc32gucbwi	chelsea masih belum konsisten dalam meraih hasil bagus di atas lapangan manajer chelsea mauricio pochettino menyebut para pemainnya merasakan tekanan
jude bellingham tak sepenuhnya fit usai mengalami cedera bahu bulan lalu. pelatih real madrid, carlo ancelotti, tak merisaukan kondisi pemainnya tersebut. https://t.co/13xm0vpqam	jude bellingham tak sepenuhnya fit usai mengalami cedera bahu bulan lalu pelatih real madrid carlo ancelotti tak merisaukan kondisi pemainnya tersebut

c. Tokenization

Dalam tahapan ini dilakukan pemisahan kata berdasarkan *whitespace.* Tahapan ini disebut dengan tokenisasi atau *tokenizing.* Tahap ini memerlukan pustaka NLTK pada python. Hasil proses ditampilkan pada tabel dibawah.

Tabel 3. Hasil tokenisasi data

Sebelum *Tokenization*	Sesudah *Tokenization*
chelsea masih belum konsisten dalam meraih hasil bagus di atas lapangan manajer chelsea mauricio pochettino menyebut para pemainnya merasakan tekanan	[chelsea, masih, belum, konsisten, dalam, meraih, hasil, bagus, di, atas, lapangan, manajer, chelsea, mauricio, pochettino, menyebut, para, pemainnya, merasakan, tekanan]
jude bellingham tak sepenuhnya fit usai mengalami cedera bahu bulan lalu pelatih real madrid carlo ancelotti tak merisaukan kondisi pemainnya tersebut	[jude, bellingham, tak, sepenuhnya, fit, usai, mengalami, cedera, bahu, bulan, lalu, pelatih, real, madrid, carlo, ancelotti, tak, merisaukan, kondisi, pemainnya, tersebut]

d. Stopwords Removal

Penghapusan *stopwords* merupakan suatu proses untuk menghapus kata-kata yang tidak relevan dan tidak memiliki makna dalam analisis data teks. Dalam proses ini, kita dapat memanfaatkan pustaka dari Python seperti NLTK dan Sastrawi (untuk kamus Bahasa Indonesia) untuk melaksanakan tugas tersebut.

Tabel 4. Hasil penghapusan *stopwords*

Sebelum *Stopwords*	Sesudah *Stopwords*
[chelsea, masih, belum, konsisten, dalam, meraih, hasil, bagus, di, atas, lapangan, manajer, chelsea, mauricio, pochettino, menyebut, para, pemainnya, merasakan, tekanan]	[chelsea, konsisten, meraih, hasil, bagus, atas, lapangan, manajer, chelsea, mauricio, pochettino, menyebut, pemainnya, merasakan, tekanan]
[jude, bellingham, tak, sepenuhnya, fit, usai,	[jude, bellingham, tak, sepenuhnya, fit, usai,

mengalami, cedera, bahu, bulan, lalu, pelatih, real, madrid, carlo, ancelotti, tak, merisaukan, kondisi, pemainnya, tersebut]	mengalami, cedera, bahu, bulan, lalu, pelatih, real, madrid, carlo, ancelotti, tak, merisaukan, kondisi, pemainnya, tersebut]

e. Stemming

Stemming adalah tahapan dalam *preprocessing* untuk menemukan kata dasar dari suatu kata yang mempunyai awalan ataupun akhiran tanpa mempertimbangkan apakah kata tersebut mempunyai makna yang sama dengan yang lainnya. Tahapan ini memerlukan pustaka Sastrawi sebagai pustaka untuk pemrosesan data teks berbahasa Indonesia.

Tabel 5. Hasil penghapusan *stemming*

Sebelum ***Stemming***	Sesudah ***Stemming***
[chelsea, konsisten, meraih, hasil, bagus, atas, lapangan, manajer, chelsea, mauricio, pochettino, menyebut, pemainnya, merasakan, tekanan]	[chelsea, konsisten, raih, hasil, bagus, atas, lapang, manajer, chelsea, mauricio, pochettino, sebut, main, rasa, tekan]
[jude, bellingham, tak, sepenuhnya, fit, usai, mengalami, cedera, bahu, bulan, lalu, pelatih, real, madrid, carlo, ancelotti, tak, merisaukan, kondisi, pemainnya, tersebut]	[jude, bellingham, tak, sepenuh, fit, usai, alami, cedera, bahu, bulan, lalu, latih, real, madrid, carlo, ancelotti, tak, risau, kondisi, main, sebut]

f. *Stopwords Removal* 2

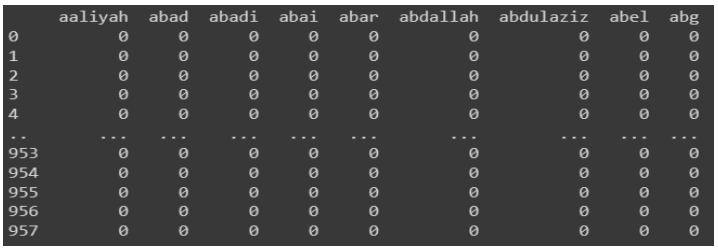
Pada tahapan *stopwords removal* kedua ini diberikan kata tambahan yaitu "detikcom", "tak", "sebut", "apa", "gt", "via", "jadi", "tinggal", "sini", "yuk", "dalam", "ten", "hag", "nih", "viral", "guna", "kata", "va", "sp", "ar", "aku", "kan", "hayo", "siapa", "nih", "yaa", "ialah", "hm", "orang", "cek", "lho", "detikers", "viral", "duh", "warganer", "salfok", "iya", "kak", "vs", "si", "lha", "kok", "kena", "tren", "kamu", "tau", "ajak", "ala", "gak", "begini", "ini", "ikut", "heboh", "duga", "situ", "sih", "banget", "keren", "kalau", "nah", "kisah", "waduh", "lalu", "bakal", "godok", "wih", "dah", "tuh", "ungkap", "halo", "bikin", "balas", "ogah", "sosok", "wah", "duga", "geger", "cuma", "kabar", "awas", "kini", "hayo", "sempat", "selamat", "hah", "enggak", "hinga", "ungkap", "cuss", "atur", "temu", "kali", "pilih", "hasil", "banyak", "soal", "rp", "atas", "hingga", "baru", "laku", "kasus", "edu", "usai", "hari", "tahun", "buat", "ri", "menang", "terus", "intip", "potret", "klub". Kata-kata tersebut dihapus karena dirasa tidak memiliki makna dan dikhawatirkan dapat mempengaruhi hasil dari pemodelan topik.

Tabel 6. Hasil penghapusan *stopwords*

Sebelum ***Stopwords 2***	Sesudah ***Stopwords 2***
[chelsea, konsisten, raih, hasil, bagus, atas, lapang, manajer, chelsea, mauricio, pochettino, sebut, main, rasa, tekan]	[chelsea, konsisten, raih, bagus, lapang, manajer, chelsea, mauricio, pochettino, main, rasa, tekan]
[jude, bellingham, tak, sepenuh, fit, usai, alami, cedera, bahu, bulan, lalu, latih, real, madrid, carlo, ancelotti, tak, risau, kondisi, main, sebut]	[jude, bellingham, sepenuh, fit, alami, cedera, bahu, bulan, latih, real, madrid, carlo, ancelotti, risau, kondisi, main]

C. Bag of Words

Hasil *pre-processing* data berupa term atau matriks yang mencakup kata-kata yang muncul secara berulang. Model *Bag of Words* digunakan untuk menghitung frekuensi kemunculan setiap kata pada term atau matriks tersebut. Hasil perhitungan jumlah kemunculan kata-kata tersebut digunakan dalam perhitungan distribusi pada metode LDA.



	aaliyah	abad	abadi	abai	abar	abdallah	abdulaziz	abel	abg
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
..	...	...	...	...	...	...	...	...	...
953	0	0	0	0	0	0	0	0	0
954	0	0	0	0	0	0	0	0	0
955	0	0	0	0	0	0	0	0	0
956	0	0	0	0	0	0	0	0	0
957	0	0	0	0	0	0	0	0	0

Gambar 2. Hasi proses *Bag of Words*

D. Analisis Pemodelan

Tahap pemodelan merupakan tahapan yang paling penting. Pemodelan dilakukan dengan menggunakan algoritma LDA dan LSA. Masing-masing model dilakukan dengan menggunakan jumlah topik, iterasi dan random state yang sama. Masing-masing model dilakukan percobaan dengan jumlah topik sebanyak 8, jumlah iterasi 10, dan jumlah *random state* 42. Hasil pemodelan kemudian diidentifikasi penentuan topik berdasarkan hasil setiap *keyword*nya yang ada didalamnya. Hasil pemodelan dijabarkan dalam tabel dibawah ini.

Tabel 7. Hasil pemodelan topik LDA dan LSA

Model	No.	Kata Kunci/*Keywords*	Topik
LDA	1	['anak', 'bunuh', 'tewas', 'polisi', 'korban', 'ayah']	Pembunuhan anak oleh ayah
	2	['tumpang', 'bom', 'canda', 'air', 'indonesia', 'pelita']	Kasus *prank* bom
	3	['tumpang', 'bom', 'canda', 'air', 'pelita', 'juanda']	Kasus *prank* bom
	4	['israel', 'gaza', 'serang', 'unru', 'gunung', 'yayu']	Perang Israel
	5	['gunung', 'marapi', 'daki', 'korban', 'polisi', 'erupsi']	Erupsi Gunung Marapi
	6	['unru', 'yayu', 'jantung', 'gunung', 'aktor', 'indonesia']	Aktor Yayu Unru meninggal dunia
	7	['ade', 'armando', 'polisi', 'jokowi', 'psi', 'lapor']	Kasus Ade Armando
	8	['city', 'manchester', 'liga', 'inggris', 'villa', 'aston']	Sepak bola
LSA	1	['ingat', 'stiker', 'tiket', 'bawaslu', 'pisa', 'tumpang', 'siswa', 'transj', 'tempel', 'dunia']	Sanksi kasus ketua Bawaslu
	2	['presiden', 'ganjar', 'gubernur', 'ruu', 'jakarta', 'ungsi', 'daftar', 'tunjuk', 'dkj', 'rumah']	RUU DKJ
	3	['israel', 'gaza', 'hamas', 'serang', 'hubner', 'indonesia', 'justin', 'jalur', 'rumah', 'warga']	Perang Israel
	4	['gunung', 'marapi', 'indonesia', 'ade', 'korban', 'daki', 'armando', 'erupsi', 'tumpang', 'psi']	Erupsi Gunung Marapi
	5	['anak', 'bunuh', 'tewas', 'ayah', 'empat', 'korban', 'jaksel', 'keluarga', 'bocah', 'jejer']	Pembunuhan anak oleh ayah

	6	['jokowi', 'bunuh', 'anak', 'panca', 'alung', 'eddy', 'wulan', 'tahan', 'kontra', 'polisi']	Pembunuhan Wulan
	7	['manchester', 'unru', 'yayu', 'city', 'covid', 'united', 'liga', 'inggris', 'chelsea', 'villa']	Sepak bola
	8	['tewas', 'polisi', 'sangka', 'satu', 'anak', 'teman', 'periksa', 'jalan', 'gratifikasi', 'kpk']	Pembunuhan anak oleh ayah

Dari hasil pemodelan kemudian diidentifikasi topik yang sesuai berdasarkan kata kunci yang ada di setiap topik. Dalam model LSA ditemukan beberapa topik yaitu pembunuhan anak oleh ayah, kasus prank bom, perang Israel, erupsi Gunung Marapi, aktor Yayu Unru meninggal dunia, kasus Ade Armando, dan sepak bola. Hasil topik dari model LDA sebagai berikut sanksi kasus ketua Bawaslu, RUU DKJ, perang Israel, erupsi Gunung Marapi, pembunuhan anak oleh ayah, pembunuhan Wulan, Sepak bola.

E. Evaluasi Model.

Setelah tahap pemodelan selesai, selanjutnya dilakukan evaluasi model dengan menggunakan *topic coherence metrics. Topic coherence* atau koherensi topik merupakan metode untuk mengetahui ukuran *coherence* yang digunakan untuk mengevaluasi nilai *topic modellingnya*. Semakin tinggi skor untuk sejumlah topik tertentu, hal ini mengindikasikan bahwa untuk setiap topik, terdapat lebih banyak kata yang saling terkait sehingga topik tersebut lebih memiliki keterkaitan yang masuk akal [21]. Oleh karena itu, untuk membandingkan algoritma pemodelan, pengukuran koherensi merupakan cara yang baik untuk digunakan.

Untuk mengukur kualitas dari model topik seperti yang dilakukan dalam penelitian ini menggunakan LSA dan LDA, dapat menggunakan metrik koherensi topik. Metrik ini mengukur seberapa sering kata-kata dalam suatu topik muncul bersama dalam korpus referensi, yang umumnya adalah dokumen-dokumen yang berkaitan dengan topik itu. Nilai koherensi topik yang tinggi menunjukkan bahwa topik itu mudah dipahami oleh manusia. Hasil nilai koherensi topi pada model LSA sebesar 0.5. Sedangkan untuk hasil dari nilai koherensi topik pada model LDA sebesar 0.45.

F. Perbandingan Performa

Perbandingan performa antara *Latent Semantic Analysis* (LSA) dan *Latent Dirichlet Allocation* (LDA) dalam topik pemodelan seringkali menjadi perdebatan yang menarik. Dalam penelitian ini, hasil nilai koherensi topik menunjukkan bahwa LSA memiliki nilai koherensi sebesar 0.5, sedangkan LDA memiliki nilai koherensi sebesar 0.45.

LSA merupakan teknik pemrosesan bahasa alami yang menggunakan metode aljabar linier untuk menemukan pola tersembunyi dalam teks. LSA cenderung lebih sederhana dalam konsepnya dan mampu menangkap hubungan sinonim antar kata serta konsep yang terkait dalam dokumen. Dalam beberapa kasus, LSA mampu memberikan hasil yang baik dalam mengidentifikasi topik dari koleksi dokumen.

Di sisi lain, LDA adalah model generatif yang menggunakan pendekatan probabilitas untuk menemukan topik tersembunyi dalam dokumen. LDA memiliki interpretabilitas yang baik karena dapat menghasilkan distribusi probabilitas kata-kata dalam setiap topik. Meskipun LDA dapat memberikan hasil yang kuat dalam mengidentifikasi topik, terkadang interpretasi topik yang dihasilkan bisa menjadi lebih rumit karena modelnya yang kompleks. Dari hasil koherensi topik yang dihasilkan, meskipun perbedaannya tidak signifikan, LSA menunjukkan nilai koherensi yang sedikit lebih tinggi (0.5) daripada LDA (0.45). Namun, keputusan terkait pilihan antara LSA dan LDA dalam topik pemodelan dapat dipengaruhi oleh berbagai faktor, seperti jenis dataset, interpretabilitas hasil, dan tujuan analisis yang diinginkan. Keduanya memiliki kelebihan dan kelemahan masing-masing, sehingga pilihan antara LSA dan LDA seringkali bergantung pada konteks spesifik dari analisis yang dilakukan.

Kesimpulan

Proses penerapan *topic modelling* dengan membandingkan model LSA dan LDA pada data *tweet* berita pada akun Detikcom diawali dengan proses *crawling data.* Proses tersebut dilakukan dengan rentang kurun waktu 1 tahun mulai dari tanggal 9 Desember 2022 sampai 9 Desember 2023 dan menghasilkan data sebanyak 958 baris. Data tersebut selanjutnya dilakukan pra-pemrosesan data (*preprocessing*) seperti *case folding*, *tokenization*,

stopwords, dan *stemming.* Kemudian hasilnya dilakukan proses *bag of words* untuk dihitung frekuensi kemunculan kata tiap dokumen. Jumlah ukuran frekuensi kemunculan tiap kata dijadikan acuan dalam pembuatan model LSA dan LDA. Masing-masing model ditentukan jumlah topik sebanyak 8, iterasi 10, dan random state sebanyak 42. Hasil pemodelan kemudian diidentifikasi topiknya berdasarkan *keywords* yang ada didalamnya.

Evaluasi dari kedua model tersebut dilakukan dengan melakukan pengukuran koherensi topik atau topic coherence menggunakan c_v. Model LSA memiliki nilai koherensi sebesar 0.5, sedangkan untuk model LDA memiliki nilai koherensi sebesar 0.45. Dalam kasus ini model LSA memiliki performa yang lebih baik daripada model LDA berdasarkan hasil nilai koherensi tiap model. Dalam pengembangan penelitian selanjutnya diharapkan peneliti dapat menggunakan kasus lain untuk pemodelan topik. Selain itu peneliti juga dapat menggunakan model lain dari pemodelan topik (*topic modelling*) seperti OCTIS dan sebagainya.

Referensi

[1] A. N. Ulfah and M. K. Anam, "Analisis Sentimen Hate Speech Pada Portal Berita Online Menggunakan Support Vector Machine (SVM)," vol. 7, no. 1, pp. 1–10, 2020, [Online]. Available: http://jurnal.mdp.ac.id

[2] C. Naury, D. H. Fudholi, and A. F. Hidayatullah, "Topic Modelling pada Sentimen Terhadap Headline Berita Online Berbahasa Indonesia Menggunakan LDA dan LSTM," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 1, p. 24, Jan. 2021, doi: 10.30865/mib.v5i1.2556.

[3] D. M. Wonohadidjojo, "Perbandingan Convolutional Neural Network pada Transfer Learning Method untuk Mengklasifikasikan Sel Darah Putih," *Ultimatics : Jurnal Teknik Informatika*, vol. 13, no. 1, p. 51, 2021.

[4] A. P. Giovani, A. Ardiansyah, T. Haryanti, L. Kurniawati, and W. Gata, "ANALISIS SENTIMEN APLIKASI RUANG GURU DI TWITTER MENGGUNAKAN ALGORITMA KLASIFIKASI," *Jurnal Teknoinfo*, vol. 14, no. 2, p. 115, Jul. 2020, doi: 10.33365/jti.v14i2.679.

[5] J. Budiarto, "Identifikasi Kebutuhan Masyarakat Nusa Tenggara Barat pada Pandemi Covid-19 di Media Sosial dengan Metode Crawling (Requirements Identification for NTB People in pandemic covid-19 at Social Media Using Crawling Method)," vol. 2, no. 4, pp. 244–250, 2021.

[6] I. N. Husada, E. H. Fernando, H. Sagala, A. E. Budiman, and H. Toba, "Ekstraksi dan Analisis Produk di Marketplace Secara Otomatis dengan Memanfaatkan Teknologi Web Crawling," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 5, no. 3, Jan. 2020, doi: 10.28932/jutisi.v5i3.1977.

[7] M. Dwirizqy Wimbassa, T. Marsyah Noor, S. Yasara, and T.

Muhammad Arsyah, “Emotional Text Detection dengan Long Short Term Memory (LSTM),” *Jurnal Format*, vol. 12, 2023.

[8] B. Gunawan, H. P. Sasty, and E. P. Esyudha, “Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes,” *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. 4, no. 2, pp. 17–29, 2018, [Online]. Available: www.femaledaily.com

[9] Samsir, Ambiyar, U. Verawardina, F. Edi, and R. Watrianthos, “Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes,” *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 1, pp. 157–163, Jan. 2021, doi: 10.30865/mib.v5i1.2604.

[10] D. Alita and A. Rahman, “Pendeteksian Sarkasme pada Proses Analisis Sentimen Menggunakan Random Forest Classifier,” 2020.

[11] M. Fiqri and R. Setya Perdana, “Klasifikasi Data Twitter pada Masa Transisi Pandemi menuju Endemi menggunakan Latent Semantic Analysis (LSA),” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 6, pp. 2736–2742, 2023, [Online]. Available: http://j-ptiik.ub.ac.id

[12] N. Hendrastuty, A. Rahman Isnain, and A. Yanti Rahmadhani, “Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine,” *Jurnal Informatika: Jurnal pengembangan IT (JPIT)*, vol. 6, no. 3, 2021, [Online]. Available: http://situs.com

Ahmad Kemal Al Izzi

Universitas Islam Negeri Sunan Ampel Surabaya

Fakultas Sains dan Teknologi

Jl. Dr. Ir. H. Soekarno No.682 Gn. Anyar Kec. Gn. Anyar, Surabaya, Jawa Timur

Rakadian Audiga Pratama

Universitas Islam Negeri Sunan Ampel Surabaya

Fakultas Sains dan Teknologi

Jl. Dr. Ir. H. Soekarno No.682 Gn. Anyar Kec. Gn. Anyar, Surabaya, Jawa Timur