DOI: 10.37014/medpus.v31i3.5237

media pustakawan

Text Mining dengan *Topic Modelling* LDA dari Pertanyaan Gelar Wicara Literasi Perpustakaan Nasional RI

Mutia Jelita

Perpustakaan Nasional Republik Indonesia, Jakarta, Indonesia
Jl. Salemba Raya No. 28A Jakarta Pusat

Korespondensi: *miu.kyu@gmail.com*

Diajukan: 27-07-2024; **Direvisi:** 03-11-2024; **Diterima:** 16-12-2024

Abstract

In 2023, the National Library of Indonesia, through its Library Analysis and Reading Culture Development Centre, organised several literacy talk shows. Each event was documented in minutes (.doc or .pdf format) recording the speakers' material, questions, and answers. As events increased, so did the volume of minutes. This research aimed to identify frequently discussed topics using Text Mining with a Topic Modelling approach. Latent Dirichlet Allocation was applied and evaluated by perplexity values (a measure of model quality). Results showed the optimal number of topics to represent the dataset was three, with the lowest perplexity value of 470.922 at the 30th iteration. The three main topics identified were reading interest and the need for books in schools and regions, libraries' role in improving children's literacy, and librarians' role in inclusive literacy programmes for both young and old, including health literacy. Frequent words were literacy, library, reading, books, and children.

Keywords: text mining; topic modelling; latent dirichlet allocation (lda); literacy talk show questions

Abstrak

Tahun 2023, Perpustakaan Nasional Republik Indonesia melalui Pusat Analisis Perpustakaan dan Pengembangan Budaya Baca menyelenggarakan berbagai gelar wicara literasi. Setiap acara didokumentasikan dalam bentuk notula berformat (.doc atau .pdf), mencatat materi narasumber, pertanyaan peserta, dan jawaban narasumber. Seiring bertambahnya acara, jumlah notula meningkat. Penelitian ini bertujuan untuk mengidentifikasi topik yang banyak dibahas menggunakan *Text Mining* dengan pendekatan *Topic Modelling*. Metode yang diterapkan adalah *Latent Dirichlet Allocation* (LDA), yang dievaluasi berdasarkan nilai *perplexity* (ukuran untuk menilai kualitas model). Hasilnya menunjukkan bahwa jumlah topik yang paling sesuai untuk menggambarkan seluruh kumpulan pertanyaan adalah tiga, dengan nilai *perplexity* terendah sebesar 470,922 pada iterasi ke-30. Tiga topik utama yang ditemukan adalah minat membaca dan kebutuhan buku di sekolah dan daerah, peran perpustakaan dalam meningkatkan literasi anak-anak, serta peran pustakawan dalam program literasi yang inklusif bagi orang muda dan tua, termasuk literasi kesehatan. Kata-kata yang paling sering muncul adalah literasi, perpustakaan, membaca, buku, dan anak.

Kata Kunci: text mining; topic modelling; latent dirichlet allocation (lda); pertanyaan gelar wicara literasi

Pendahuluan

Penggunaan informasi yang tepat dapat meningkatkan efisiensi kerja dalam berbagai bidang, termasuk sektor pemerintahan. Ilmu informasi yang muncul sejak tahun 1950 dan masih berkembang sampai sekarang merupakan ilmu interdisipliner yang membahas tentang interpretasi informasi dan berbagai hal terkait dengan proses kognitif, konteks, dan fenomena informasi (Priyanto, 2013). Ilmu ini mengajarkan bagaimana mengelola dan memanfaatkan data secara efektif.

Artificial Intelligence (AI) atau kecerdasan buatan kini menjadi bagian penting dari perkembangan ilmu informasi, termasuk untuk analisis data. Rich & Knight (1991, dikutip dari Soyusiawaty, 2023) menyebutkan bahwa AI merupakan sebuah studi tentang bagaimana membuat komputer melakukan hal-hal yang pada saat ini dapat dilakukan lebih baik oleh manusia. Beberapa cabang ilmu sub-bagian dari AI, yakni *Expert System* (Sistem Pakar), *Speech Understanding* (Pemahaman Ucapan), *Robotics and Sensory System* (Robotik dan Sensor Sistem), *Pattern Recognition* (Pengenalan Pola), dan *Natural Language Processing* (Soyusiawaty, 2023). Dengan bantuan AI, pengolahan informasi dalam jumlah besar dapat dilakukan dengan cepat dan akurat.

NLP merupakan salah satu teknik yang berfungsi untuk melakukan *text mining*. Salah satu pendekatan dalam *text mining*, yaitu *Topic Modelling.* Pendekatan ini tergolong dalam metode *machine learning* yang digunakan untuk menemukan "topik" yang muncul dalam kumpulan dokumen. Selain itu, pendekatan ini juga tergolong dalam metode pembelajaran tanpa pengawasan (*unsupervised learning*), sehingga tidak harus memutuskan sebelumnya topik apa yang akan dijadikan target. Singkat kata, *topic modelling* dapat digunakan untuk menemukan *underlying structure* 'struktur yang mendasari' dalam suatu teks (Silge, 2017).

Sepanjang tahun 2023, Perpustakaan Nasional Republik Indonesia (Perpusnas) melalui salah satu unit kerjanya yakni Pusat Analisis Perpustakaan dan Pengembangan Budaya Baca (PAPPBB) telah banyak mengadakan gelar wicara literasi. Acara ini menjadi bagian dari kegiatan Peningkatan Indeks Literasi Masyarakat (PILM), Duta Baca Indonesia (DBI), dan Sosialisasi Pembudayaan Kegemaran Membaca melalui Webinar yang diselenggarakan di berbagai kota di Indonesia baik secara luring maupun daring. Biasanya setelah diawali dengan paparan materi para narasumber, kemudian kegiatan dilanjutkan dengan sesi tanya-jawab.

Catatan mengenai jalannya kegiatan, hal yang dibahas, dan pertanyaan-pertanyaan dari para peserta dalam kegiatan gelar wicara tersebut, lalu dikumpulkan dalam suatu notula dengan format dokumen (.doc atau .pdf). Notula ini dari waktu ke waktu jumlahnya bertambah banyak, begitu juga dengan pertanyaan-pertanyaan yang terkumpul. Bila dari sebagian kegiatan literasi selama satu tahun di satu unit kerja saja sudah menghasilkan data sebegitu banyak, dapatlah dibayangkan melimpahnya data dari seluruh kegiatan segenap unit kerja selama bertahun-tahun. Dari notula-notula ini dapat digali informasi lebih dalam, termasuk topik dalam pertanyaan yang sering diajukan oleh para peserta dan rangkuman materi para narasumber. Suatu langkah manual tentu bukan hal yang tepat dilakukan di tengah perkembangan teknologi informasi saat ini.

Di sinilah *text mining* dengan pendekatan *topic modelling* dapat memainkan peran dan urgensinya, sehingga menjadi hal yang menarik bagi penulis untuk diteliti. Pemodelan topik dapat menangkap isi yang tersembunyi dari kumpulan data berupa teks yang jumlahnya sangat banyak. Dengan menggunakan pendekatan *topic modelling,* penulis ingin memahami apa saja topik-topik yang sering diperbincangkan serta kata-kata apa yang paling banyak muncul di dalam gelar wicara literasi yang diselenggarakan Perpusnas tersebut. Temuan penelitian ini diharapkan dapat menjadi fokus prioritas untuk perencanaan program atau kebijakan yang hendak direncanakan untuk tahun-tahun mendatang.

Tinjauan Pustaka

Gelar Wicara Literasi

Dalam rangka meningkatkan kecerdasan kehidupan bangsa, Perpusnas selaku Lembaga Pemerintah Non Departemen (LPND) yang salah satu fungsinya sebagai perpustakaan pembina menurut Undang-Undang Nomor 43 Tahun 2007 tentang perpustakaan, perlu menumbuhkan budaya gemar membaca melalui pengembangan dan pendayagunaan perpustakaan. Salah satu hal yang dilakukan untuk mendukung hal itu adalah dengan menyelenggarakan beberapa kegiatan, yakni Peningkatan Indeks Literasi Masyarakat (PILM), Duta Baca Indonesia (DBI) dan Sosialisasi

Pembudayaan Kegemaran Membaca melalui Webinar. Ketiga kegiatan tersebut memiliki tema yang berbeda-beda di setiap acara tetapi dua diantaranya memiliki tema besar, yaitu (a) PILM: Peningkatan Indeks Literasi Masyarakat untuk Kesejahteraan dan (b) DBI: Membaca itu Sehat, Menulis itu Hebat.

Kegiatan-kegiatan tersebut memiliki tujuan, diantaranya meningkatkan promosi tentang peran perpustakaan dan budaya literasi terhadap kesejahteraan masyarakat, meningkatkan kesadaran masyarakat untuk gemar membaca dan memanfaatkan perpustakaan sebagai sarana pembelajar sepanjang hayat, mendorong apresiasi kepada insan maupun kelompok masyarakat yang berkontribusi aktif dalam pemberdayaan literasi di lingkungannya serta bagi para pemangku kepentingan untuk berperan dalam penguatan budaya literasi.

Gelar wicara literasi merupakan salah satu sesi dengan format bincang-bincang yang terdapat dalam susunan kegiatan. Gelar wicara adalah perbincangan yang dipandu oleh pembawa acara dan dihadiri oleh narasumber ahli dalam bidang tertentu (Ridwan & Azizah, 2022). Sedangkan literasi menurut Eisner yaitu kemampuan menangkap makna dari bentuk representasi yang ada di sekitar, bukan hanya simbol konvensional berupa tulisan. (Abidin et al., 2021).

Tetapi pada kegiatan yang dilaksanakan oleh Perpusnas ini, diskusi atau tanya-jawab yang terjadi adalah antara para narasumber, seperti pimpinan daerah, anggota legislatif, akademisi, dan pegiat literasi dengan peserta yang terdiri dari Kepala OPD, kepala sekolah, guru, pustakawan, penulis, mahasiswa, siswa, pengelola taman bacaan/perpustakaan komunitas, dan sebagainya dipandu oleh pembawa acara. Para narasumber memaparkan materi masing-masing kemudian dilanjutkan dengan peserta yang mengajukan pertanyaan. Pertanyaan-pertanyaan dan jawaban-jawaban dari gelar wicara literasi yang diselenggarakan di berbagai daerah di Indonesia ini dikumpulkan dalam suatu notula.

Text Mining

Menurut pengertiannya, *text mining* dipahami sebagai penggunaan teknik AI dan NLP untuk mengekstraksi informasi dan wawasan dari teks. Teknik-teknik ini mengubah data yang tidak terstruktur menjadi data terstruktur sehingga memudahkan dalam analisis data (*Text Mining vs. NLP*, 2023). Tujuan *text mining* adalah untuk pengkategorian data teks, baik itu untuk menemukan kategori yang sesuai dengan kelas yang ditentukan (*supervised learning*) maupun berdasarkan kesamaan karakteristik dan memberikan label pada kelas yang belum diketahui (*unsupervised learning*) (Rahayu et al., 2024). Dalam kaitan dengan dataset pertanyaan-pertanyaan, *text mining* pernah dilakukan oleh Rohim et al., (2023) untuk meningkatkan kemampuan *chatbot QnA (Questions and Answers)* dalam memahami pertanyaan pelanggan dan memberikan jawaban yang akurat. Hasil penelitian yang diterapkan untuk PT. PLN (Persero) Sumatera Selatan ini adalah alat bantu berupa aplikasi tanya-jawab informasi yang interaktif layaknya model diskusi dan dapat menggunakan bahasa sehari-hari.

Topic Modelling

Topic Modelling merupakan salah satu pendekatan *text mining* yang berguna untuk mengidentifikasi topik-topik yang muncul dalam suatu data teks berdasarkan kemiripannya tanpa harus memiliki label/kategori sebelumnya (Narendra, 2022; Zahra & Carkiman, 2024). Beberapa contoh algoritma *topic modelling* yang umum dipakai adalah *Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), dan Latent Semantic Analysis (LSA)* (Polin, M, et al, 2022).

Dalam penelitiannya, Silge (2017) menggunakan *topic modelling* untuk mengetahui topik-topik apa saja yang ada pada dataset pertanyaan dari *Questions and Answers (QnA) website* Stack Overflow mengenai *programming*. Dari kumpulan pertanyaan tersebut, digali topik yang sering ditanyakan oleh para pengunjung *website*. Hasil penelitian menunjukkan 12 kluster topik, diantaranya *front-end web development, databases, C and low-level programming*, dan sebagainya.

Pendekatan ini menggunakan model statistik untuk menemukan topik dalam dataset teks, kata mana yang berkontribusi ke suatu topik, dan topik mana yang berkontribusi pada suatu dokumen. Oleh karena itu, *topic modelling* sangat berguna saat kita ingin memahami isi yang mendominasi dalam kumpulan dokumen yang berjumlah sangat banyak, tanpa harus membaca semua teks secara manual.

Latent Dirichlet Allocation (LDA)

Metode LDA adalah model statistik yang mencoba menangkap topik tersembunyi dalam kumpulan dokumen (Hidayat, et al, 2015). Hidayat et al. menerapkan metode LDA dalam meringkas teks secara otomatis untuk meng-*cluster* 398 dataset dari *blog* publik. Sohrabi et al., (2017) juga mengimplementasikan LDA untuk mengklasifikasikan artikel mengenai dunia maya (*cyberspace*). Dalam penelitian tersebut, terdapat 5 topik yang terbentuk dari 1.860 artikel. Kata-kata dan frasa kunci seperti "*platform*", "IoT", "*cloud*", "inovasi" "*Cyber*", "sistem", "keamanan", "data", "perangkat lunak", dan "jaringan" sering digunakan dalam artikel.

Kelebihan dari metode ini adalah dapat meringkas, meng-*cluster*, menghubungkan atau memproses data yang sangat besar serta menghasilkan daftar topik yang diberi bobot untuk masing-masing dokumen (Chilmi, 2021). Menurut Blei (2003, dikutip dari Putra & Kusumawardani, 2017), ide dasar dari metode LDA yaitu setiap dokumen direpresentasikan sebagai campuran acak atas topik yang tersembunyi, dimana setiap topik memiliki karakter yang ditentukan berdasarkan distribusi kata-kata yang terdapat didalamnya.

LDA juga digunakan oleh Narendra (2022) dalam penelitiannya untuk menentukan nama *intent* dalam pembuatan *chatbot* yang telah memiliki data percakapan dibandingkan membuat data pelatihan percakapan dari awal sehingga lebih efektif. Untuk menentukan jumlah topik terbaik dilakukan perhitungan pada *coherence score* pada jumlah n topik.

Text Preprocessing

Sebelum dianalisis menggunakan metode LDA, dilakukan *text preprocessing*, yakni suatu proses pengubahan bentuk kata yang belum terstruktur menjadi data yang terstruktur sesuai dengan kebutuhan untuk proses *mining* yang lebih lanjut (*sentiment analysis*, peringkasan, *clustering* dokumen, dsb.) (Soyusiawaty, 2023). Beberapa *text preprocessing* yang umumnya diterapkan, yaitu:

1. *Parsing*
   Pemecahan struktur dokumen menjadi komponen-komponen terpisah (Soyusiawaty, 2023, hlm. 19).
2. *Case Folding*
   Menurut Bagus (2017, dikutip dari Chilmi, 2021), *case folding* bertujuan untuk menghindari adanya dua kata yang sama namun dianggap berbeda karena perbedaan huruf kapital dan huruf kecil.
3. *Lexical Analysis/Tokenization*
   Memisahkan setiap kata yang menyusun suatu dokumen dengan melakukan penghilangan angka, tanda baca dan karakter selain huruf alfabet, karena karakter-karakter tersebut dianggap sebagai pemisah kata (*delimiter*) dan tidak memiliki pengaruh terhadap pemrosesan teks atau menggunakan teknik pemisahan kata yang lebih canggih seperti pemisahan berbasis aturan atau pemisahan berbasis pembelajaran mesin supaya menghasilkan elemen yang bermakna (Soyusiawaty, 2023; Zahra dan Carkirman, 2024; Verma, et al, 2014).
4. *Stopword Removal/Filtering*
   Pemilihan kata-kata penting dari hasil tokenisasi, yaitu kata-kata apa saja yang akan digunakan untuk mewakili dokumen atau penghapusan *stopwords* (kosa kata yang bukan merupakan ciri

(kata unik) dari suatu dokumen) yang dapat dibuang dengan pendekatan *bag-of-words* untuk meningkatkan performa sistem.

5. *Bag-of-words*
   Representasi numerik dari dokumen sebagai vektor kata-kata yang muncul. Setiap kata-kata yang muncul ini dapat diberi bobot berdasarkan metode tertentu, contohnya TF-IDF (*Term Frequency-Inverse Document Frequency*). Teknik TF-IDF pernah digunakan Matira et al. (2023) untuk pembobotan kata dalam memodelkan topik pada judul berita *online* detik.com. Pada penelitian tersebut dihasilkan 3 topik, yaitu topik I: membahas konflik dan krisis negara; topik II: mengenai isu berkaitan dengan krisi kemanusian dan topik III: tentang isu korupsi oleh pejabat negara. Penelitian ini dilakukan dengan dataset yang diambil pada periode 16 Desember 2021 hingga 24 Maret 2022.

Metode Penelitian

Penulis menggunakan penerapan *text mining* dengan metode *Latent Dirichlet Allocation* (LDA) untuk melihat topik-topik apa saja yang sering ditanyakan oleh para peserta gelar wicara dan kata-kata yang sering muncul dalam pertanyaan-pertanyaan tersebut. Terdapat empat tahapan yang dilakukan dalam penelitian ini, yakni sebagai berikut:

Tahap I: Formulasi Masalah

Masalah atau ruang lingkup yang dibahas dalam penelitian ini adalah ingin mengetahui topik apa saja yang sering ditanyakan peserta gelar wicara literasi Perpustakaan Nasional RI dan kata-kata apa saja yang sering muncul sehingga topik-topik tersebut bisa digunakan sebagai bahan pertimbangan dalam menentukan program atau tema kegiatan di tahun mendatang.

Tahap II: Pengumpulan Data

Penulis mendapatkan data dari himpunan notula berbagai kegiatan dalam Google Drive di unit kerja PAPPBB, di mana penulis bekerja, yang hanya bisa diakses oleh pegawai PAPPBB, seperti PILM, DBI dan Sosialisasi Pembudayaan Kegemaran Membaca melalui Webinar dengan format dokumen .doc dan .pdf. Tahun 2023 PAPPBB telah berhasil menyelenggarakan 87 kegiatan berformat gelar wicara dengan jumlah detail tiap kegiatan, yaitu: (a) PILM 62 kegiatan; (b) DBI 12 kegiatan; (c) Sosialisasi Pembudayaan Kegemaran Membaca 13 kegiatan.

Dari 87 kegiatan itu, hanya ada 39 notula yang mencatat pertanyaan dari peserta. Jumlah pertanyaan dari setiap kegiatan berbeda-beda, rata-rata 3 pertanyaan. Setelah menghimpun notula, penulis melakukan *parsing*, yaitu memilah pertanyaan dari dokumen notula dan mengumpulkannya menjadi *file* berformat .xlsx. Data yang terkumpul berjumlah 145 pertanyaan.

Tahap III: Pemilihan dan Penerapan Pendekatan *Text Mining*

Penulis memilih *Latent Dirichlet Allocation (LDA)* sebagai metode yang akan diterapkan dalam melakukan *topic modeling* untuk dataset pertanyaan gelar wicara. LDA dapat mengidentifikasi topik-topik yang sering muncul dalam dataset pertanyaan dengan melihat pola dan menghubungkannya sehingga menjadi kelompok-kelompok teks.

Sedangkan *tools* yang dipakai untuk melakukan proses LDA adalah RapidMiner yang sekarang bernama Altair AI Studio. Penulis mengaplikasikan Altair AI Studio versi 2024.0.3. Perangkat lunak ini tidak berbayar dan menggunakan bahasa pemrograman Java sehingga dapat digunakan dalam berbagai macam Sistem Operasi.

Tahap IV: Evaluasi Hasil

Setelah dilakukan analisis menggunakan LDA, dilakukan evaluasi terhadap hasilnya dengan melihat nilai *perplexity* (nilai kerancuan). Nilai *perplexity* menentukan jumlah iterasi metode yang digunakan dengan cara menjalankan pemodelan terhadap topik dengan setidaknya tiga parameter topik yang berbeda (Putra & Kusumawardani, 2017). Semakin rendah nilai *perplexity* maka jumlah topik yang dihasilkan akan semakin baik.

Hasil dan Pembahasan

Pemrosesan Teks (*Text Preprocessing*)

Seperti telah dijelaskan pada bagian Metode Penelitian, di tahap pengumpulan data dilakukan *parsing* untuk mengambil pertanyaan-pertanyaan dari notula. Setelah *parsing*, dataset pertanyaan dalam format Excel diimpor ke Altair AI Studio selanjutnya dilakukan *text preprocessing* lain berupa *case folding* dan *tokenization*.

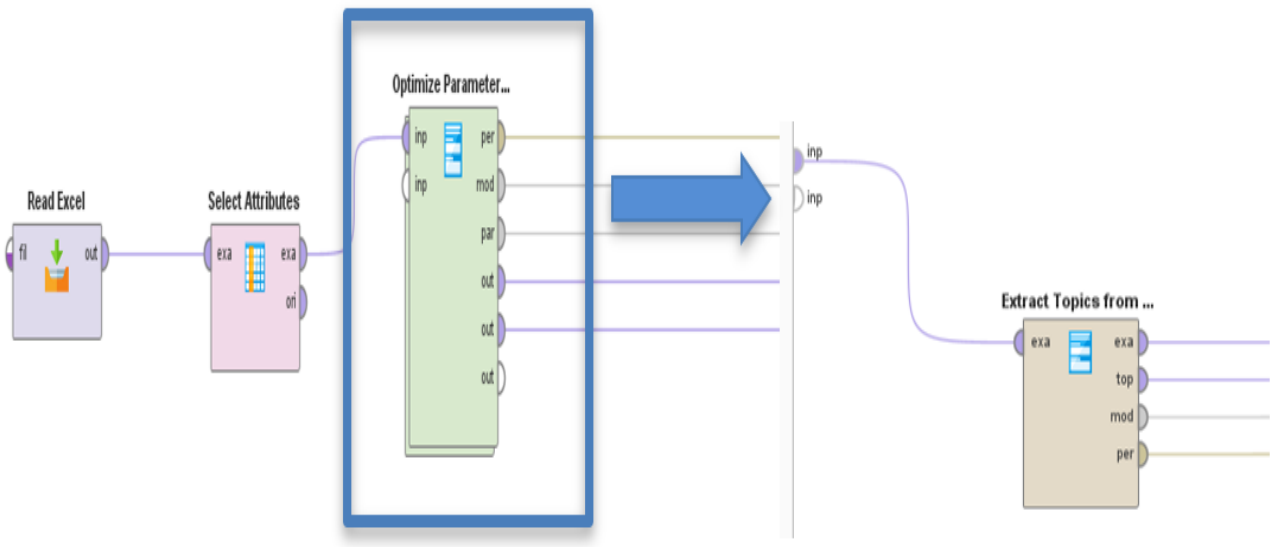
Kata-kata dalam dataset pertanyaan gelar wicara literasi masih ada yang berbentuk huruf besar dan kecil, ini akan berpengaruh pada saat proses analisis *topic modelling* karena kata yang sama tetapi yang satu dalam huruf besar dan lainnya huruf kecil akan dianggap berbeda oleh operator LDA. Sehingga harus disamakan semua bentuk hurufnya menjadi huruf kecil saja atau huruf besar saja. Dalam hal ini, penulis memilih agar semua kata dalam dataset pertanyaan menjadi bentuk huruf kecil.

Saat semua kata telah menjadi huruf kecil, kemudian dilakukan *tokenization,* yaitu memisahkan setiap kata yang menyusun suatu pertanyaan. Selain itu menghilangkan angka, tanda baca dan karakter selain huruf alfabet, karena karakter-karakter tersebut tidak akan berpengaruh dalam proses analisis *topic modelling*. Dalam dataset pertanyaan gelar wicara literasi, terdapat beberapa kata dalam bahasa asing, yakni Bahasa Inggris. Ini dapat mempengaruhi hasil pemrosesan teks sehingga harus diseragamkan terlebih dahulu menjadi satu bahasa yang sama. Tetapi karena keterbatasan waktu, penulis tidak melakukan hal tersebut.

Setelah itu, *bag-of-words* diterapkan pada setiap kata yang telah terpisah untuk diberikan bobot berdasarkan seberapa sering kata tersebut muncul dalam sebuah dokumen (*term frequency*) dan seberapa umum atau jarang kata tersebut muncul di seluruh dokumen dalam korpus (*inverse document frequency*). Dalam daset pertanyaan gelar wicara literasi, yang digunakan hanya *term frequency (TF)*. Bobot diberikan berdasarkan seberapa sering kata muncul dalam keseluruhan dataset pertanyaan. Tidak seperti pemrosesan teks lainnya yang memiliki operator tersendiri, *bag-of-words* akan diproses dalam operator LDA sekaligus untuk dilakukan proses analisa. Ini karena LDA fokus pada pola *co-occurrence* (kemunculan bersama) kata dalam dokumen untuk menemukan topik. Hasil dari pemrosesan teks ini masih berupa file *Excel*.

Proses Analisis *Topic Modelling*

Proses selanjutnya adalah melakukan analisis *topic modelling* dengan operator LDA seperti ditunjukkan Gambar 1 di bawah ini:



Gambar 1. Proses Topic Modelling dengan metode LDA

Terlihat beberapa operator, yaitu *Read Excel* (membaca dataset dalam bentuk Excel), *Select Attributes* (memilih atribut/variabel dalam tabel yang akan diproses), *Optimize Parameter* (mengoptimalkan parameter), dan *Extract Topics From Data (LDA)* (memproses *topic modelling*). Operator LDA disematkan di dalam operator *Optimize Parameter* yang memiliki gambaran berbeda dari operator lainnya, yakni seperti *file* bertumpuk. *File Excel* hasil pemrosesan teks, dibaca (*read*) menggunakan operator *Read Excel* kemudian dipilih atribut atau nama kolom yang akan digunakan pada proses *topic modelling*.

Pada operator LDA ada parameter yang harus diisi sesuai kebutuhan, seperti jumlah topik, jumlah iterasi, banyaknya kata dalam 1 topik, dan jenis *stopword*. Dalam penelitian ini penulis memasukkan jumlah topik sebanyak 10 dan iterasi 100. Tidak ada aturan baku untuk memasukkan jumlah topik dan iterasi karena sebanyak apapun akan tetap dioptimalkan oleh operator *Optimize Parameters.*

Pada setiap iterasi, LDA memperbarui pengelompokan kata dalam topik untuk mendekati hasil yang paling representatif. Melalui proses ini, model mengidentifikasi pola distribusi kata dengan lebih baik. LDA menggunakan iterasi untuk mencapai titik di mana perubahan dalam pengelompokan topik menjadi stabil. Ini menandakan bahwa model sudah menemukan struktur topik yang optimal. Selanjutnya, menentukan banyaknya kata dalam 1 topik, penulis memasukkan angka 5. Sehingga nantinya akan ada 5 kata yang menyusun suatu topik tertentu. Kemudian dari kelima kata tersebut bisa diinterpretasikan topiknya.

Setelah itu, memilih jenis *stopwords. Stopword language* yang harus dipilih adalah Bahasa Indonesia karena dataset pertanyaan yang menjadi bahan penelitian berbahasa Indonesia walaupun ada beberapa kata dalam Bahasa Inggris. *Stopword* ini merupakan kumpulan kata dalam bahasa Indonesia yang akan dibuang atau tidak dipakai, misalnya kata “bahwa”, “jika”, “selalu” dan lain-lain. Altair AI Studio telah menyediakan pilihan “Bahasa Indonesia” untuk *stopword*-nya sehingga tidak perlu mengunggah *library* berisi kumpulan kata yang akan dibuang tersebut, misalnya seperti *library* Sastrawi.

Evaluasi Hasil Proses Analisis *Topic Modelling*

Hasil proses analisis *text mining* dengan pendekatan *topic modelling* menggunakan metode *Latent Dirichlet Allocation* (LDA) ini tampak pada Tabel 1 berikut:

Tabel 1. Hasil analisis LDA (jumlah topik dan *perplexity*)

Iteration	Extract Topic from Data (LDA) (2).number_of_topics	Extract Topic from Data (LDA) (2).iterations	Perplexity
1	3	5	472,103
2	7	5	474,756
3	3	10	471,830
4	7	10	473,774
5	3	15	472,361
6	7	15	472,310
7	3	20	472,682
8	7	20	474,666
9	3	25	471,198
10	7	25	474,294
11	3	30	470,922
12	7	30	471,838

Pada Tabel 1 di atas, menunjukan bahwa nilai *perplexity* terendah ada pada saat jumlah topik 3 dan iterasinya sebanyak 30 iterasi dengan nilai *perplexity* 470,922 (berwarna kuning). Sehingga dari dataset pertanyaan gelar wicara literasi, terbentuk jumlah topik yang paling tepat menggambarkan keseluruhan dataset pertanyaan gelar wicara literasi, yaitu 3 topik. Dengan semakin banyaknya iterasi, LDA berusaha mengurangi nilai *perplexity* yang menunjukkan seberapa baik model memprediksi topik pada data.

Masing-masing pertanyaan memiliki nilai *confidence* yang menunjukkan hubungan yang paling tinggi dengan masing-masing topik. Kebalikan dari nilai *perplexity*, nilai *confidence* akan semakin baik jika nilainya semakin besar. Untuk lebih detail, dapat dilihat pada Gambar 2.

Contohnya, pertanyaan pertama, pertanyaan ini diprediksi masuk ke dalam topik ketiga (*Topic_2*) karena memiliki nilai *confidence* yang tinggi pada topik ketiga (*Topic_2*), yaitu 0,351. Lain halnya dengan pertanyaan nomor 13, pertanyaan tersebut masuk dalam topik kedua (*Topic_1*) dengan nilai *confidence* 0,348. Kemudian pertanyaan nomor 6 yang mempunyai nilai *confidence* 0,405, sehingga masuk ke topik pertama (*Topic_0*).

Jumlah pertanyaan yang masuk ke dalam masing-masing topik pertama sampai ketiga secara berurutan adalah 51 pertanyaan, 51 pertanyaan, dan 43 pertanyaan. Demikian, *Topic_0* dan *Topic_1* memiliki jumlah anggota pertanyaan yang sama.

Open in Turbo Prep Auto Model Interactive Analysis Filter (145 / 145 examples): all

Row No.	documentid	prediction(Topic)	confidence(Topic_0)	confidence(Topic_1)	confidence(Topic_2)	Questions
1	0	Topic_2	0.315	0.333	0.351	ada 1 tbm yang mengelola bakau, apakah perpusnas bisa bekerja sama lebih aktif dengan tbm?
2	1	Topic_0	0.363	0.319	0.319	ada kasus pasangan suka berolahraga dengan frekuensi tinggi tetapi ada masalah dalam funsi seksualitas. up
3	2	Topic_2	0.321	0.330	0.348	adakah cara kerjasama dengan mahasiswa untuk open rekruitmen
4	3	Topic_2	0.333	0.324	0.342	adakah efek samping vitamin pada anak-anak dalam jangka panjang?
5	4	Topic_1	0.333	0.356	0.310	adakah faktor-faktor yang membuat sulawesi tenggara tertinggal dari daerah lain dalam hal literasi. literasi tidak
6	5	Topic_1	0.308	0.350	0.342	adakah kebijakan untuk mengekspresikan karya agar dapat dimanfaatkan masyarakat untuk dapat meningkatk
7	6	Topic_0	0.405	0.323	0.271	adanya pendefinisian bahwa proses membaca buku yang benar itu harus membaca buku tertentu terlebih dah
8	7	Topic_1	0.327	0.345	0.327	alasan apa yang diambilnya warna hitam putih untuk photo story?
9	8	Topic_2	0.301	0.341	0.358	apa gerakan nyata yang dilakukan perpustakaan terhadap program transformasi perpustakaan berbasis inklusi
10	9	Topic_2	0.324	0.324	0.352	apa hal yang dilakukan ketika berada dititik terendah dalam karier?
11	10	Topic_1	0.327	0.337	0.337	apa itu marhaen?
12	11	Topic_2	0.313	0.339	0.348	apa pembelajaran bais dari perang ini yang harus dilaporkanke panglima? melihat di negara sendiri seperti kas
13	12	Topic_1	0.330	0.348	0.322	apa saja strategi dan program untuk meningkatkan literasi di daerah terpencil kabupaten enrekang
14	13	Topic_1	0.315	0.369	0.315	apa sektor paling krusial dari sektor literasi dan ekonomi kreatif?
15	14	Topic_1	0.327	0.346	0.327	apa trik untuk membuat anak-anak gemar membaca
16	15	Topic_1	0.327	0.337	0.337	apa yang harus dilakukan saat stuck.
17	16	Topic_2	0.339	0.313	0.348	apakah ada bantuan yang bersubsidi untuk penambahan buku koleksi diperpustakaan desa atau di tbm

ExampleSet (145 examples,5 special attributes,1 regular attribute)

Gambar 2. Detail hasil analisis LDA (pembagian topik per pertanyaan)

Interpretasi Topik

Berdasarkan hasil prediksi LDA, kata-kata dalam dataset pertanyaan gelar wicara literasi, terdistribusi ke dalam 3 topik. Kata-kata tersebut memiliki nilai yang koheren satu sama lain dalam suatu topik. Penulis telah menentukan jumlah kata untuk ditampilkan pada setiap topik, yaitu 5 seperti tampak pada Gambar 3.

LDAPrediction

```
LDA Model with 3 topics
alphaSum = 50.0
beta = 0.06993006993006994
Topic 0 tokens=415.0000 document_entropy=4.6167 word-length=5.8000      coherence=-32.0950      uniform_dist=1.5236     corpus_dist=1.0490      eff_num_words=87.7356
  membaca       word-length=7.0000      coherence=0.0000        uniform_dist=0.1927     corpus_dist=0.0591      token-doc-diff=0.0000   exclusivity=0.9938
  buku  word-length=4.0000      coherence=-2.1667       uniform_dist=0.1279     corpus_dist=0.0325      token-doc-diff=0.0000   exclusivity=0.7519
  daerah        word-length=6.0000      coherence=-5.5545       uniform_dist=0.0876     corpus_dist=0.0323      token-doc-diff=0.0003   exclusivity=0.9888
  sekolah       word-length=7.0000      coherence=-5.2306       uniform_dist=0.0780     corpus_dist=0.0296      token-doc-diff=0.0000   exclusivity=0.9878
  minat word-length=5.0000      coherence=-5.0645       uniform_dist=0.0686     corpus_dist=0.0225      token-doc-diff=0.0004   exclusivity=0.8204
Topic 1 tokens=451.0000 document_entropy=4.6648 word-length=9.0000      coherence=-31.2170      uniform_dist=1.5540     corpus_dist=0.9909      eff_num_words=60.7348
  literasi      word-length=8.0000      coherence=0.0000        uniform_dist=0.3340     corpus_dist=0.0847      token-doc-diff=0.0010   exclusivity=0.9959
  perpustakaan  word-length=12.0000     coherence=-2.4068       uniform_dist=0.2355     corpus_dist=0.0641      token-doc-diff=0.0058   exclusivity=0.9946
  anak  word-length=4.0000      coherence=-5.3043       uniform_dist=0.1054     corpus_dist=0.0343      token-doc-diff=0.0000   exclusivity=0.9899
  anak-anak     word-length=9.0000      coherence=-2.5764       uniform_dist=0.0784     corpus_dist=0.0275      token-doc-diff=0.0003   exclusivity=0.9875
  meningkatkan  word-length=12.0000     coherence=-5.3043       uniform_dist=0.0697     corpus_dist=0.0252      token-doc-diff=0.0008   exclusivity=0.9863
Topic 2 tokens=400.0000 document_entropy=4.6728 word-length=5.8000      coherence=-37.5789      uniform_dist=1.2413     corpus_dist=1.0558      eff_num_words=156.5558
  lari  word-length=4.0000      coherence=0.0000        uniform_dist=0.0721     corpus_dist=0.0288      token-doc-diff=0.0048   exclusivity=0.9872
  program       word-length=7.0000      coherence=-4.2836       uniform_dist=0.0625     corpus_dist=0.0259      token-doc-diff=0.0007   exclusivity=0.9858
  orang word-length=5.0000      coherence=-4.7484       uniform_dist=0.0442     corpus_dist=0.0202      token-doc-diff=0.0003   exclusivity=0.9819
  pustakawan    word-length=10.0000     coherence=-4.7484       uniform_dist=0.0356     corpus_dist=0.0173      token-doc-diff=0.0002   exclusivity=0.9790
  tua   word-length=3.0000      coherence=-4.7484       uniform_dist=0.0356     corpus_dist=0.0173      token-doc-diff=0.0002   exclusivity=0.9790
```

Gambar 3. Kata-kata dalam tiap topik

Setiap topik memiliki kata-kata utama, beserta informasi tambahan seperti panjang kata, nilai koherensi (*coherence*), eksklusivitas (*exclusivity*), entropi dokumen (document_entropy), dan efektivitas jumlah kata (*eff_num_words*). Berikut ini interpretasi penulis untuk masing-masing topik:

1. *Topic_0* yang merupakan topik pertama terdiri dari 5 kata utama berikut: membaca, buku, daerah, sekolah, dan minat. Kata "membaca" memiliki koherensi 0,0000 (paling relevan), sedangkan "daerah" memiliki koherensi -5,5545 (kurang relevan). Eksklusivitas kata berkisar antara 0,7519 (kata "buku") hingga 0,9938 (kata "membaca"). Topik pertama ini memiliki 415 token dan *eff_num_words* sebesar 87,7356. Dari kata-kata tersebut, *Topic_0* menunjukkan bahasan mengenai minat membaca dan kebutuhan buku di sekolah dan daerah.
2. *Topic_1* terdiri dari 5 kata utama berikut: literasi, perpustakaan, anak, anak-anak, dan meningkatkan. Kata "literasi" memiliki koherensi 0,0000 (paling relevan), sementara "anak" memiliki koherensi -5,3043 (kurang relevan). Eksklusivitas berkisar antara 0,9863 (kata "meningkatkan") hingga 0,9959 (kata "literasi"). Topik kedua mempunyai jumlah token sebanyak 451 token dan *eff_num_words* sebesar 60,7348. Dengan kumpulan kata-kata tersebut, dapat dirangkai bahwa *Topic_1* mengenai peran perpustakaan dalam meningkatkan literasi anak-anak.
3. *Topic_2* terdiri dari kumpulan 5 kata utama berikut: lari, program, orang, pustakawan, dan tua. Kata "lari" memiliki koherensi 0,0000 (paling relevan), sedangkan "program" memiliki koherensi -4,2836 (kurang relevan). Eksklusivitas berkisar antara 0,9790 (kata "pustakawan" dan "tua") hingga 0,9872 (kata "lari"). Terdapat 400 token pada *Topic_2* dan *eff_num_words* sebesar 156,5558. Penulis menyimpulkan dari kata-kata dalam *Topic_2* membahas mengenai peran pustakawan dalam program literasi yang inklusif bagi orang muda dan tua, termasuk literasi kesehatan.

Pada ketiga topik yang terbentuk, nilai koheren yang dihasilkan kurang relevan karena bernilai negatif atau menjauhi nilai nol (*Topic_0*: -32,0950; *Topic_1*: -31,2170; dan *Topic_2*: -37,5789). Koherensi dapat mengukur seberapa "bermakna" dan "terkait" kata-kata dalam sebuah topik. Rentang spesifik nilai koherensi sangat bergantung pada metode perhitungan dan data yang digunakan. Semakin besar nilainya, semakin baik kualitas topiknya. Penelitian ini menunjukkan bahwa kata-kata dalam topik kurang terhubung secara semantik atau memiliki kualitas yang beragam atau jarang muncul bersama kata-kata utama dalam dataset pertanyaan gelar wicara literasi.

Kemudian ada nilai eksklusivitas, nilai ini menggambarkan seberapa unik kata tersebut untuk topik tertentu. Kata-kata dengan eksklusivitas tinggi lebih spesifik untuk topik tertentu dan tidak

sering muncul di topik lainnya. Misalnya, kata "literasi" memiliki eksklusivitas 0,9959 di *Topic_1*, menunjukkan bahwa kata ini hampir sepenuhnya unik untuk topik tersebut. Nilai eksklusivitas lebih rendah seperti pada kata “buku” di *Topic_0* (0,7519) menunjukkan bahwa kata tersebut mungkin muncul di lebih dari satu topik, sehingga kurang unik untuk satu topik saja. Semua kata dalam model memiliki eksklusivitas tinggi (di atas 0,75), menunjukkan bahwa kata-kata ini cukup spesifik untuk topik masing-masing. Ini adalah tanda yang baik karena menunjukkan adanya pembagian kata yang relatif unik antara topik.

Hal lain yang perlu diperhatikan adalah keragaman distribusi topik berdasarkan nilai entropi dokumen. Semakin tinggi entropi (menjauhi 0) maka semakin besar keragaman distribusi topik dalam dokumen. Artinya, dokumen tersebut membahas banyak topik secara hampir merata, tanpa dominasi satu topik tertentu. Sebaliknya, semakin rendah entropi (mendekati 0), semakin terfokus dokumen tersebut pada satu topik tertentu. Ini menunjukkan dokumen tersebut sangat spesifik dan berkonsentrasi pada satu topik utama. Dalam LDA, perhitungan nilai entropi dokumen bergantung pada jumlah topik. Pada penelitian ini, entropi dokumen bernilai rata-rata 4,6. Ini mengindikasikan tingkat variasi topik yang cukup besar dalam dataset pertanyaan gelar wicara literasi.

Selain itu, setiap topik memiliki jumlah kata efektif yang berkontribusi secara signifikan dalam mendeskripsikan sebuah topik pada dokumen terkait, dalam hal ini dataset pertanyaan gelar wicara literasi. Jika nilainya tinggi maka topik memiliki banyak kata yang relevan dan tersebar. Sedangkan nilai rendah menandakan topik lebih terfokus pada sedikit kata yang dominan. Dalam penelitian ini, rentang keseluruhan nilai, yaitu 60,7348 hingga 156,5558. Topik yang lebih fokus (*Topic_1*) lebih mudah diinterpretasikan karena kata-katanya lebih terarah. Topik yang lebih luas (*Topic_2*) mencakup lebih banyak kata relevan sehingga memiliki variasi kata yang lebih besar.

Kata yang Sering Muncul

Selanjutnya, hasil proses LDA juga menyatakan kata-kata yang sering muncul dari 145 pertanyaan yang menjadi sampel penelitian. Hal ini dihitung dengan melihat bobot masing-masing kata. Lima kata teratas yang memiliki bobot tertinggi atau yang paling sering muncul, yaitu literasi (37 kali), perpustakaan (28 kali), membaca (22 kali), buku (16 kali), dan anak (15 kali). Tabel 2 di bawah ini, memperlihatkan bobot dari 15 kata teratas.

Tabel 2. Kata-kata yang sering muncul dalam dataset pertanyaan

topic.id	word	weight
1	literasi	37
1	perpustakaan	28
0	membaca	22
0	buku	16
1	anak	15
0	daerah	12
1	anak-anak	12
0	sekolah	11
1	meningkatkan	11
0	minat	10
2	lari	10
2	program	9
2	orang	7
2	pustakawan	6
2	tua	6

Kata-kata yang sering muncul seperti dalam tabel di atas dapat juga disajikan dalam visualisasi agar terlihat lebih jelas dan menarik, misalnya dengan menggunakan *wordcloud* 'awan kata' seperti pada Gambar 4.

Gambar 4. Wordcloud kata-kata yang sering muncul dalam dataset pertanyaan

Kesimpulan

Dari penelitian ini diketahui terbentuk 3 topik hasil *text mining* menggunakan pendekatan *topic modelling* dengan metode *Latent Dirichlet Allocation (LDA)*. Nilai ini dapat diidentifikasi dengan melihat nilai *perplexity* terendah, yaitu 470.922, yang ada pada iterasi ke-30. Ketiga topik tersebut adalah (1) minat membaca dan kebutuhan bantuan buku di sekolah dan daerah; (2) peran perpustakaan dalam meningkatkan literasi di kalangan anak-anak; serta (3) peran pustakawan dalam program literasi yang inklusif bagi orang muda dan tua, termasuk literasi kesehatan. Sedangkan kata-kata yang paling sering muncul dari 145 dataset pertanyaan, adalah literasi (37 kali), perpustakaan (28 kali), membaca (22 kali), buku (16 kali), dan anak (15 kali).

Dari temuan di atas, *text mining* terbukti dapat menjadi solusi untuk menggali isi dari berbagai informasi maupun wacana yang melimpah dalam dokumen yang jumlahnya sangat banyak. Dengan pendekatan *topic modelling,* dapat diketahui topik-topik apa saja yang sering dibicarakan dalam kegiatan gelar wicara literasi yang diselenggarakan Perpusnas tahun 2023. Hasil ini dapat digunakan sebagai fokus prioritas untuk perencanaan program kegiatan atau kebijakan yang hendak direncanakan untuk tahun-tahun mendatang sehingga dapat dipilih topik-topik tertentu yang sekiranya perlu penguatan, atau bisa juga untuk meminimalkan pengulangan, sehingga topik yang disajikan ke depan dapat lebih beragam, baik secara meluas maupun mendalam.

Studi lebih lanjut dapat dikembangkan dengan menerapkan *text preprocessing* lanjutan seperti *phrase detection* (mendeteksi adanya 2 kata atau lebih yang menjadi frasa), *stemming* (membuat kata menjadi kata dasar), dan penyeragamkan bahasa atau menggunakan *stopwords* untuk multibahasa sehingga analisis kata menjadi lebih akurat. Selain itu, dengan menggunakan *text mining* dari data tanya-jawab gelar wicara literasi, dapat direkomendasikan pula penyusunan daftar pertanyaan yang sering diajukan (FAQ).

Daftar Pustaka

Abidin, Y., et al (2021). *Pembelajaran literasi: Strategi meningkatkan kemampuan literasi matematika, sains, membaca dan menulis*. Bumi Aksara.

Chilmi, M. L. C. (2021). *Latent dirichlet allocation (LDA) untuk mengetahui topik pembicaraan warganet twitter tentang omnibus law* [Skripsi, Universitas Islam Negeri Syarif Hidayatullah]. Institutional Repository UIN Syarif Hidayatullah. https://repository.uinjkt.ac.id/dspace/bitstream/123456789/56724/1/M.%20LUVIAN%20CHISNI%20CHILMI-FST.pdf

Hidayat, E., et al (2015). Automatic text summarization using latent dirichlet allocation (LDA) for document clustering. *International Journal of Advances in Intelligent Informatics, 1*(3), 132-139. https://doi.org/10.26555/ijain.v1i3.43

Indonesia. (2007). *Undang-undang republik indonesia nomor 43 tahun 2007 tentang perpustakaan*. Lembaran Negara Republik Indonesia Tahun 2007 Nomor 129.

Matira, et al. (2023). Pemodelan topik pada judul berita online detikcom menggunakan latent dirichlet allocation. *Estimasi: Journal of Statistics and Its Application, 4*(1), 53-63. https://doi.org/10.20956/ejsa.vi.24843

Narendra, L. W. (2022). Topic modeling in conversational dialogs for naming intent labels using lda. *Jurnal Sistem Telekomunikasi, Elektronika, Sistem Kontrol, Power System & Komputer, 2*(1), 65-74. https://doi.org/10.32503/jtecs.v2i1.1820

Polin, M., et al. (2022). Analisa dan visualisasi hasil kuesioner pertanyaan terbuka menggunakan elasticsearch dan kibana. *Jurnal Sistem Informasi (E-Journal), 14*(2), 2763–2777. https://doi.org/10.18495/jsi.v14i2.18418

Priyanto, I. F. (2013). Apa dan mengapa ilmu informasi. *Jurnal Kajian Informasi & Perpustakaan, 1*(1), 55-59. https://doi.org/10.24198/jkip.v1i1.9611

Putra, I. M. K. B & Kusumawardani, R. P. (2017). Analisis topik informasi publik media sosial di Surabaya menggunakan pemodelan latent dirichlet allocation (LDA). *Jurnal Teknik ITS, 6*(2), A311-A316. *https://*doi: 10.12962/j23373539.v6i2.23205

Rahayu, et al. (2024). *Buku ajar data mining*. PT. Sonpedia Publishing Indonesia.

Ridwan, M. H. & Azizah, L., (2022). Analisis struktur percakapan Merry Riana dan narasumber pada gelar wicara "zero to hero". *PENEROKA: Jurnal Kajian Ilmu Pendidikan Bahasa dan Sastra Indonesia, 2*(1), 67-80. https://doi.org/10.30739/peneroka.v2i1.1366

Rohim, A., et al. (2023). Penerapan metode *text mining* dengan *chatbot questions and answers* pada PT. PLN (persero) sumatera selatan. *Klik-Jurnal Ilmu Komputer 4*(2), 59-67. https://doi.org/10.56869/klik.v4i2.551

Silge, Julia. (2017). *Text mining of stack overflow questions*. Stackoverflow. https://stackoverflow.blog/2017/07/06/text-mining-stack-overflow-questions/.

Sohrabi, B., Raeesi Vanani, I., & Baranizade Shineh, M. (2018). Topic modeling and classification of cyberspace papers using text mining. *Journal of Cyberspace Studies, 2*(1), 103-125. https://doi.org/10.22059/jcss.2017.239847.1009

Suyosiawaty, Dewi. (2023). *Buku ajar pemrosesan bahasa alami*. Laboratorium teknik Informatika Universitas Ahmad Dahlan.

Text mining vs. NLP: What's the difference?. (2023, June 20). Relative Insight. https://relativeinsight.com/text-mining-vs-nlp/

Verma, T. et al. (2014). Tokenization and filtering process in rapidminer. *International Journal of Applied Information Systems (IJAIS), 7*(2), 16-18. Foundation of Computer Science Inc. https://doi.org/10.5120/ijais14-451139

Zahra, D. F. & Carkirman. (2024). Pengalaman pelanggan membeli tiket konser coldplay: Menambang ulasan online berdasarkan pemodelan topik dan analisis sentimen. *Journal of Information System, Applied, Management, Accounting and Research, 8*(2), 243-260. https://doi.org/10.52362/jisamar.v8i2.1426