SPIRIT VOL 16 NO 1 (2024) 170 - 182
JURNAL SPIRIT
225/E/KPT/2022

e-ISSN 2721-057X
p-ISSN 2085-3092

Pencarian Topik Penelitian Pada Studi Kasus Jurnal JIFTI Menggunakan Teknik *Hiearchical Dirichlet Processes*

Moch Erreza [1], Kartini[2], Agung Mustika Rizki[3]
[1,2,3] Universitas Pembangunan Nasional "Veteran" Jawa Timur
mo.erreza091@gmail.com[1], kartini.if@upnjatim.ac.id[2], agung.mustika.if@upnjatim.ac.id[3]

Naskah diterima: 24 Pebruari 2024 ; Direvisi : 16 Mei 2024 ; Disetujui : 16 Mei 2024

Abstrak

Studi ini mengusulkan proses *Dirichlet* hierarkis (HDP), karena kami melaporkan hasil pengujian pada tiga set dokumen yang menunjukkan kinerja HDP yang unggul dan efisien dibandingkan dengan model sebelumnya. Oleh karena itu, penelitian ini menggunakan metode HDP. Objek dalam penelitian ini adalah pemodelan topik/*topic modelling* pada dokumen Jurnal Teknologi Informasi dan Robotika (JIFTI). Penelitian ini menggunakan data abstrak Jurnal Teknologi Informasi dan Robotika (JIFTI) dari tahun 2019 sampai 2022, yang diperoleh dari https://jifti.upnjatim.ac.id/index.php/jifti/issue/archive. Hasil dari pre-processing kemudian dihitung dengan menggunakan *topic modeling Hierarchical Dirichlet Process* (HDP) untuk melihat topic 20 dengan melihat kata yang sering muncul pada data abstrak Jurnal Teknologi Informasi dan Robotika (JIFTI). Jumlah kemunculan setiap kata tersebut menjadi ukuran dalam metode *Hierarchical Dirichlet Process* (HDP) untuk dimodelkan *Topic modeling* abstrak Jurnal Teknologi Informasi dan Robotika (JIFTI) menggunakan metode *Hierarchical Dirichlet Process* (HDP) akan diketahui yang paling banyak muncul, Temuan utama meliputi sejumlah kata yang sering muncul, seperti "*practicum*, *test*, *manage*, *major*, *develop*, *aim*, *feed*, *bitching*, *technology*, *people*, *apply*, *tourism*, *student*, *user*, *feed*, *learn*, *digit*, *laboratory*, *pusvetma*, *product*". Dari 20 kata yang sering muncul dalam setiap topik, dapat dilihat bahwa mayoritas abstrak jurnal JIFTI menyoroti penelitian dalam implementasi.

Kata kunci: Topik Penelitian, HDP dan *Text Preprocessing*

Abstract

This study proposes a hierarchical Dirichlet process (HDP), as we report test results on three sets of documents that demonstrate the superior and efficient performance of HDP compared to previous models. Therefore, this research uses the HDP method. The object of this research is topic modeling/topic modeling in Journal of Information Technology and Robotics (JIFTI) documents. This research uses abstract data from the Journal of Information Technology and Robotics (JIFTI) from 2019 to 2022, obtained from https://jifti.upnjatim.ac.id/index.php/jifti/issue/archive. The results of pre-processing are then calculated using Hierarchical Dirichlet Process (HDP) topic modeling to view 20 topics by looking at words that frequently appear in the Journal of Information Technology and Robotics (JIFTI) abstract data. The number of occurrences of each word is a measure in the Hierarchical Dirichlet Process (HDP) method to be modeled. Journal of Information Technology and Robotics (JIFTI) abstract modeling topics using the Hierarchical Dirichlet Process (HDP) method will determine which ones appear the most. The main findings include a number of frequently used words. Appears, such as "practicum, test, manage, major, develop, aim, feed, grumble, technology, people, apply, tourism, student, user, feed, learn, digit, laboratory, pusvetma, product". From the 20 words that frequently appear in each topic, it can be seen that the majority of JIFTI journal abstracts highlight research in implementation.

***Keywords**: Topic modeling, HDP dan Text Preprocessing,*

PENDAHULUAN

Dalam konteks pendidikan tinggi di Indonesia, karya ilmiah merupakan persyaratan yang harus dipenuhi oleh mahasiswa untuk mendapatkan gelar Sarjana Strata-1. Setiap perguruan tinggi memiliki ketentuan tersendiri dalam penyelesaian tugas akhir, yang dapat berupa skripsi, tesis, jurnal ilmiah, artikel, prototipe, dan berbagai bentuk lainnya. Salah satu komponen penting dalam karya ilmiah adalah abstrak, yang berfungsi sebagai ringkasan singkat dari keseluruhan isi penelitian. Abstrak ini menyajikan pokok-pokok penting dari penelitian yang kemudian akan dijelaskan secara lebih lengkap dalam isi jurnal ilmiah[1].

Berdasarkan Jurnal Ilmiah Teknologi Informasi dan Robotika yang selalu bertambah setiap tahunnya, maka seharusnya terdapat informasi dari kumpulan dokumen jurnal tersebut. Meskipun jumlah dokumen jurnal terus meningkat setiap tahun, pada umumnya, belum ada penelitian lanjutan yang dapat memberikan ringkasan informasi yang memadai dari dokumen-dokumen tersebut. Oleh karena itu, dengan pertambahan jurnal setiap tahun dan pertumbuhan jumlah dokumen jurnal, penting untuk mengadopsi metode *text mining* untuk mengeksplorasi dan mengelola informasi yang terdapat dalam kumpulan dokumen jurnal tersebut.

Saat ini, banyak perusahaan mengadopsi sistem data *mining* untuk melakukan segmentasi pasar. Data *mining* merupakan proses eksplorasi dan penemuan pengetahuan dari kumpulan data yang tersedia. Proses ini melibatkan penggunaan berbagai metode seperti analisis statistik, penerapan konsep matematika, kecerdasan buatan, dan teknik pembelajaran mesin untuk mengambil informasi dan pengetahuan yang bermanfaat dari berbagai kumpulan data besar. Hal ini memungkinkan perusahaan untuk mengoptimalkan strategi pemasaran dan mengidentifikasi peluang bisnis yang berpotensi[2][3][4].

Data *mining* tidak hanya digunakan dalam analisis perusahaan, melainkan juga dalam berbagai bidang lain. Contoh aplikasinya meliputi telekomunikasi untuk menganalisis transaksi, keuangan untuk mendeteksi transaksi mencurigakan, dan penjelajahan internet untuk menganalisis perilaku pelanggan[5][6][7]. Selain itu, terdapat bidang serupa seperti penambangan teks, yang fokus pada penemuan pola dalam data tekstual besar. Tujuannya adalah untuk menemukan informasi yang berguna untuk berbagai tujuan[8][9].

Salah satu peran dari penambangan data dan penambangan teks adalah pengelompokan (*clustering*). Pengelompokan merupakan metode dalam penambangan data atau teks yang tidak memerlukan supervisi. Metode ini tidak membutuhkan data latihan atau label target keluaran. Clustering terbagi menjadi dua kategori

utama: clustering hierarki dan clustering non-hierarki. Pendekatan hierarki melibatkan pengelompokan dua atau lebih objek yang memiliki kesamaan yang paling tinggi terlebih dahulu. Langkah ini kemudian diulangi dengan menambahkan objek tambahan ke dalam kelompok tersebut. Ini menghasilkan struktur pohon hierarki di mana hubungan antara objek diwakili sebagai tingkat (level) pada pohon. Di sisi lain, pengelompokan non-hierarkis dimulai dengan menentukan jumlah *cluster* yang diinginkan terlebih dahulu, dan kemudian objek dikelompokkan berdasarkan kriteria tertentu tanpa memperhatikan struktur hierarki[10][11][12].

Dalam Jurnal Ilmiah Teknologi Informasi dan Robotika, permasalahan yang muncul adalah belum adanya pengetahuan tentang jumlah klaster topik penelitian yang ada. Penerapan teknik *clustering* pada topik penelitian diperlukan untuk memahami tren topik yang sedang berkembang. Pengelompokan topik bisa dilakukan secara manual oleh siswa, tetapi membutuhkan waktu yang cukup lama. Untuk mengatasi kendala tersebut, penggunaan komputer dengan teknik pemodelan dapat menjadi solusi yang lebih efisien. Secara umum, terdapat dua jenis penulisan esai, salah satunya terfokus pada konteks teknologi informasi dan robotika[13].

Pemodelan topik merupakan metode pengelompokan non-hierarki yang secara otomatis mengorganisir topik yang muncul dari data untuk membentuk kelompok topik terkait. Pendekatan ini bertujuan untuk mengatasi tantangan yang sering dihadapi saat menganalisis jurnal ilmiah di bidang teknologi informasi dan robotika. Terdapat berbagai metode pemodelan topik, termasuk Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), dan Hierarchical Dirichlet Process (HDP). LSA merupakan teknik pemrosesan bahasa alami yang bertujuan untuk mengungkap makna dengan menciptakan representasi vektor teks. PLSA, yang diperkenalkan oleh Puzicha dan Hofmann, adalah pendekatan untuk mengindeks dokumen secara otomatis menggunakan model kelas statistik laten, dengan upaya untuk meningkatkan aspek probabilistik dari metode LSA[14]. LDA, yang dikembangkan oleh Blei, Ng, dan Theo, adalah peningkatan dari PLSA yang menggabungkan elemen pemodelan campuran untuk menangkap interaksi antara kata dan dokumen. HDP, di sisi lain, adalah model untuk data yang dapat dikelompokkan di mana setiap item data dapat termasuk ke dalam beberapa kelompok yang berbeda, seperti yang terjadi dalam model topik di mana kata-kata dikelompokkan ke dalam dokumen yang mungkin memiliki topik-topik yang saling tumpang tindih[15].

Dalam penelitian ini, memanfaatkan model Bayesian nonparametrik yang dikenal sebagai Hierarchical Dirichlet Process (HDP)

untuk mengelompokkan masalah dengan beberapa kumpulan data. HDP memungkinkan pemodelan setiap kelompok data sebagai gabungan komponen terbuka yang secara otomatis dihasilkan dari model tersebut. HDP memiliki kemampuan untuk memisahkan komponen di antara kelompok-kelompok tersebut, memungkinkan penggunaan pemodelan ketergantungan antar kelompok secara efisien. Hasil pengujian pada tiga set dokumen menunjukkan bahwa HDP menunjukkan kinerja yang lebih unggul dan efisien dibandingkan dengan model sebelumnya. Sehingga, HDP digunakan dalam Jurnal Ilmiah Teknologi Informasi dan Robotika (JIFTI) sebagai bagian dari penelitian ini.

METODE

Desain penelitian yang dilakukan dalam penelitian ini adalah berisi proses pencarian informasi yang berhubungan dengan pemodelan topik. Yakni meliputi proses pencarian abstrak skripsi yang berkaitan dengan *topic modeling* menggunakan metode *Hierarchical Dirichlet Process* (HDP). Pengumpulan data yang melibatkan proses pengambilan data abstrak skripsi dari situs web Jurnal Ilmiah Teknologi Informasi dan Robotika (JIFTI) (https://jifti.upnjatim.ac.id/index.php/jifti/index). Pengambilan data dilakukan menggunakan alat bantu dalam Google Chrome yang disebut sebagai *google colab*. Objek dalam penelitian ini adalah pemodelan topik/*topic modeling* pada dokumen Jurnal Ilmiah Teknologi Informasi dan Robotika (JIFTI) dari tahun 2019-2022. Metode Analisis ini melibatkan beberapa tahap, yaitu tahap *Preprocessing data* yang terdiri dari: *Casefolding, Remove Stopword, Filtering Frequent Words, Filtering Rare Words, Stemming, Normalisasi, Tokenizing, Simpanan Dataset Bersih, Topic modeling, analisis tren.*

HASIL DAN PEMBAHASAN

1. *Pre-Procesing* Data

Penelitian ini memanfaatkan platform *Google Colab* sebagai alat untuk mengatur data. Platform ini secara khusus difungsikan untuk melakukan data mining dan mengelola data dengan menggunakan pendekatan pemrograman visual. *Google Colab* menyajikan berbagai widget yang dapat digunakan untuk membangun model, dalam rangka melakukan proses pengelolaan data dan data *mining*. Tahap *Pre-processing* data terdiri dari delapan langkah. Ini melibatkan *casefolding, remove stopwords, filtering frequent words, filtering rare words, stemming,* normalisasi, *tokenizing,* dan penyimpanan dataset bersih.

a. *Casefolding*

Casefolding merupakan suatu teknik dalam pengolahan teks yang mengharmonisasikan semua karakter menjadi huruf kecil. Pada bagian ini, terdapat beberapa langkah pemrosesan yang

dilakukan untuk membersihkan dan mempersiapkan teks agar lebih mudah untuk diolah atau diambil informasinya[1].

Fungsi preprocess_text di atas bertujuan untuk melakukan pra-pemrosesan pada sebuah teks atau kalimat, dan salah satu langkah yang dilakukan adalah casefolding. Casefolding merupakan suatu teknik dalam pengolahan teks yang mengharmonisasikan semua karakter menjadi huruf kecil. Pada bagian ini, terdapat beberapa langkah pemrosesan yang dilakukan untuk membersihkan dan mempersiapkan teks agar lebih mudah untuk diolah atau diambil informasinya.

Pertama, kalimat awal diubah menjadi huruf kecil menggunakan fungsi lower(), sehingga tidak ada perbedaan huruf besar dan kecil dalam analisis teks. Selanjutnya, angka dihapus dari kalimat menggunakan ekspresi reguler re.sub(r"\d+", "", lower_case) agar fokus pada kata-kata dan bukan angka.

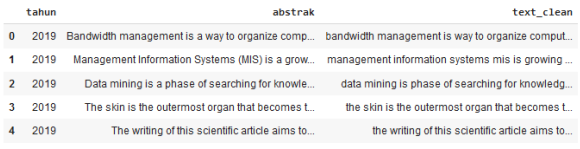
Langkah selanjutnya adalah menghapus tanda baca dari kalimat menggunakan translate dan string.punctuation. Ini membantu membersihkan teks dari tanda baca yang tidak relevan dalam analisis. Kemudian, spasi pada awal dan akhir kalimat dihilangkan dengan menggunakan strip() untuk menghindari adanya spasi yang tidak diinginkan.

Selanjutnya, URL, termasuk yang dimulai dengan "http", "https", atau "www", dihapus dari kalimat menggunakan ekspresi reguler re.sub(r"(?:\@|http?\://|https?\://|www)\S+", "", hasil). Ini penting untuk menghilangkan hyperlink yang tidak diperlukan.

Beberapa langkah lainnya melibatkan penghapusan karakter HTML, mempertimbangkan hanya huruf dan angka, mengganti karakter baru dengan spasi, dan menghapus karakter tunggal yang tidak relevan. Semua langkah ini bertujuan untuk membersihkan teks dari elemen-elemen yang tidak dibutuhkan.

Walaupun terdapat bagian yang di-comment (tidak aktif), seperti penghapusan emoji, namun pada implementasi ini bagian tersebut di-comment. Jika diperlukan, dapat diaktifkan kembali sesuai kebutuhan.

Secara keseluruhan, fungsi preprocess_text tersebut adalah langkah-langkah yang umum dilakukan dalam pra-pemrosesan teks untuk memastikan teks bersih, konsisten, dan siap untuk analisis lebih lanjut. *Output* yang dihasilkan dari tahapan *coding* di atas ialah sebagai berikut:



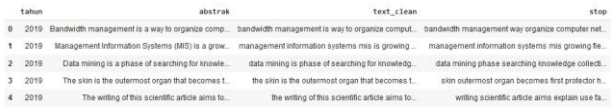
	tahun	abstrak	text_clean
0	2019	Bandwidth management is a way to organize comp...	bandwidth management is way to organize comput...
1	2019	Management Information Systems (MIS) is a grow...	management information systems mis is growing ...
2	2019	Data mining is a phase of searching for knowle...	data mining is phase of searching for knowledg...
3	2019	The skin is the outermost organ that becomes t...	the skin is the outermost organ that becomes t...
4	2019	The writing of this scientific article aims to...	the writing of this scientific article aims to...

Gambar 1. Tahap *Casefolding*
Sumber: Google colab

b. *Remove Stopwords*

Penghapusan *stopword* adalah proses di mana kata-kata "*stopword*" atau "kata pengisi" dihilangkan dari teks atau dokumen

tertentu. *Stopword* adalah kata-kata umum yang sering muncul dalam bahasa tertentu dan biasanya tidak memberikan informasi penting atau kontekstual dalam analisis teks. Menghapus stopword adalah langkah umum dalam pra-pemrosesan teks dalam pemrosesan bahasa alami (NLP) dan pengambilan informasi. Hasil dari *Output* tersebut seperti pada Gambar 2:



	tahun	abstrak	text_clean	stop
0	2019	Bandwidth management is a way to organize comp...	bandwidth management is way to organize comput...	bandwidth management way organize computer net...
1	2019	Management Information Systems (MIS) is a grow...	management information systems mis is growing ...	management information systems mis growing fie...
2	2019	Data mining is a phase of searching for knowle...	data mining is phase of searching for knowledg...	data mining phase searching knowledge collecti...
3	2019	The skin is the outermost organ that becomes t...	the skin is the outermost organ that becomes t...	skin outermost organ becomes first protector h...
4	2019	The writing of this scientific article aims to...	the writing of this scientific article aims to...	writing scientific article aims explain use fa...

Gambar 2. Remove Stopwords
(Sumber: Google colab)

Stopword dalam bahasa Inggris kode ini bertujuan untuk membersihkan teks dari stop words dalam bahasa Inggris, kemudian hasilnya akan disimpan dalam kolom baru "stop" pada DataFrame. Hal ini berguna saat melakukan pemrosesan teks lebih lanjut seperti analisis teks atau pemodelan bahasa alami.

c. *Filtering Frequent Words* dan *Rare Words*

Tujuan utama dari *Filtering Frequent Words* adalah mengurangi kebisingan dalam teks dan memfokuskan perhatian pada kata-kata kunci atau informasi yang lebih bermakna. Kata-kata yang muncul sangat sering seperti "system," "data," "information," "method," "research," and "results" dalam bahasa Inggris adalah contoh umum dari kata-kata frequent yang sering dihapus dalam proses ini. *Output* dari hasil *coding* tersebut seperti pada Gambar 3:

	tahun	abstrak	text_clean	stop	stq
0	2019	Bandwidth management is a way to organize comp...	bandwidth management is way to organize comput...	bandwidth management way organize computer net...	bandwidth management way organize computer
1	2019	Management Information Systems (MIS) is a grow...	management information systems mis is growing ...	management information systems mis growing fie...	management systems mis growing field scien
2	2019	Data mining is a phase of searching for knowle...	data mining is phase of searching for knowledg...	data mining phase searching knowledge collecti...	mining phase searching knowledge collectio
3	2019	The skin is the outermost organ that becomes t...	the skin is the outermost organ that becomes t...	skin outermost organ becomes first protector h...	skin outermost organ becomes first protect
4	2019	The writing of this scientific article aims to...	the writing of this scientific article aims to...	writing scientific article aims explain use fa...	writing scientific article aims explain us

Gambar 3. *Filtering Frequent Words*
(Sumber: Google colab)

Filtering Frequent Words membantu meningkatkan relevansi analisis teks dengan menghilangkan kata-kata yang sering kali tidak memiliki makna yang signifikan dalam konteks tertentu. Namun, seperti dalam kasus penghapusan stopword, perlu diingat bahwa tindakan ini harus disesuaikan dengan tujuan analisis. Beberapa konteks, seperti pemrosesan informasi atau analisis sentimen, beberapa kata-kata frequent mungkin memiliki makna penting, dan penghapusan dapat menghilangkan informasi yang diperlukan.

d. *Stemming*

Stemming dilakukan untuk menggabungkan kata-kata yang memiliki akhiran berbeda tetapi memiliki makna yang sama atau mirip. Ini membantu mengurangi variasi kata dalam teks sehingga analisis teks dapat lebih fokus pada makna dasar kata-kata. *Output* dari hasil *coding* tersebut seperti pada Gambar 4:

	tahun	abstrak	text_clean	stop	stopfreq	stopfreqrare	stemmed
0	2019	Bandwidth management is a way to organize comp...	bandwidth management is way to organize comput...	bandwidth management way organize computer net...	bandwidth management way organize computer net...	bandwidth management way organize computer net...	bandwidth manag way organ comput network bandw...
1	2019	Management Information Systems (MIS) is a grow...	management information systems mis is growing ...	management information systems mis growing fie...	management systems mis growing field science d...	management systems mis growing field science d...	manag system mi grow field scienc develop fiel...
2	2019	Data mining is a phase of searching for knowle...	data mining is phase of searching for knowledg...	data mining phase searching knowledge collecti...	mining phase searching knowledge collection mi...	mining phase searching knowledge collection mi...	mine phase search knowledg collect mine also e...
3	2019	The skin is the outermost organ that becomes t...	the skin is the outermost organ that becomes t...	skin outermost organ becomes first protector h...	skin outermost organ becomes first protector h...	skin outermost organ becomes first protector h...	skin outermost organ becom first protector hum...
4	2019	The writing of this scientific article aims to...	the writing of this scientific article aims to...	writing scientific article aims explain use fa...	writing scientific article aims explain use fa...	writing scientific article aims explain use fa...	write scientif articl aim explain use face det...

Gambar 4. *Stemming*
(Sumber: Google colab)

Stemming umumnya dilakukan dengan menghilangkan akhiran kata sesuai dengan aturan-aturan tertentu. Proses ini

tidak selalu sempurna, dan dalam beberapa kasus, bisa menghasilkan bentuk kata dasar yang tidak selalu valid atau bermakna. Misalnya, dalam *Stemming*, kata "*running*" dapat diubah menjadi "*run*," tetapi "*run*" mungkin tidak lagi memiliki makna yang sama dengan "*running*" dalam semua konteks.

e. *Normalisasi*

Normalisasi teks melibatkan serangkaian langkah seperti mengonversi teks ke huruf kecil (*lowercasing*), menghapus tanda baca, menghilangkan karakter khusus, menggabungkan variasi ejaan kata yang sama, dan menghapus kata-kata umum (*stopwords*). Ini membantu menciptakan representasi teks yang lebih konsisten untuk analisis lebih lanjut. *Output* dari hasil *coding* tersebut seperti pada Gambar 5:

	tahun	abstrak	text_clean	stop	stopfreq	stopfreqrare	stemmed	normal
0	2019	Bandwidth management is a way to organize comp...	bandwidth management is way to organize comput...	bandwidth management way organize computer net...	bandwidth management way organize computer net...	bandwidth management way organize computer net...	bandwidth manag way organ comput network bandw...	bandwidth manage way organ compute network ban...
1	2019	Management Information Systems (MIS) is a grow...	management information systems mis is growing ...	management information systems mis growing fie...	management systems mis growing field science d...	management systems mis growing field science d...	manag system mi grow field scienc develop fiel...	manage system mi grow field science develop fi...
2	2019	Data mining is a phase of searching for knowle...	data mining is phase of searching for knowledg...	data mining phase searching knowledge collecti...	mining phase searching knowledge collection mi...	mining phase searching knowledge collection mi...	mine phase search knowledg collect mine also e...	mine phase search knowledge collect mine also ...
3	2019	The skin is the outermost organ that becomes t...	the skin is the outermost organ that becomes t...	skin outermost organ becomes first protector h...	skin outermost organ becomes first protector h...	skin outermost organ becomes first protector h...	skin outermost organ becom first protector hum...	skin outermost organ become first protector hu...
4	2019	The writing of this scientific article aims to...	the writing of this scientific article aims to...	writing scientific article aims explain use fa...	writing scientific article aims explain use fa...	writing scientific article aims explain use fa...	write scientif articl aim explain use face det...	write scientific article aim explain use face ...

Gambar 5 Normalisasi
(Sumber: Google colab)

Normalisasi berperan penting dalam pemrosesan data dan analisis, membantu dalam membuat data atau teks menjadi lebih mudah dipahami, dibandingkan, dan digunakan dalam berbagai konteks.

f. *Tokenizing* dan Simpan *Dataset* Bersih

Tujuan dari tokenisasi adalah untuk memudahkan pemrosesan dan analisis teks, sehingga komputer dapat memahami dan memanipulasi teks dengan lebih efisien. *Output* dari hasil *coding* tersebut seperti pada Gambar 6:

	tahun	abstrak	text_clean	stop	stopfreq	stopfreqrare	stemmed	lemmatized	normal	token
0	2019	Bandwidth management is a way to organize comp...	bandwidth management is way to organize comput...	bandwidth management way organize computer net...	bandwidth management way organize computer net...	bandwidth management way organize computer net...	bandwidth manag way organ comput network bandw...	bandwidth management way organize computer net...	bandwidth manage way organ compute network ban...	['bandwidth, manage, way, organ, compute, netwo...
1	2019	Management Information Systems (MIS) is a grow...	management information systems mis is growing ...	management information systems mis growing fie...	management systems mis growing field science d...	management systems mis growing field science d...	manag system mi grow field scienc develop fiel...	management system mis grow field science devel...	manage system mi grow field science develop fi...	['manage, system, mi, grow, field, science, dev...
2	2019	Data mining is a phase of searching for knowle...	data mining is phase of searching for knowledg...	data mining phase searching knowledge collecti...	mining phase searching knowledge collection mi...	mining phase searching knowledge collection mi...	mine phase search knowledg collect mine also e...	mine phase search knowledge collect mine also ...	mine phase search knowledge collect mine also ...	['mine, phase, search, knowledge, collect, mine...
3	2019	The skin is the outermost organ that becomes t...	the skin is the outermost organ that becomes t...	skin outermost organ becomes first protector h...	skin outermost organ becomes first protector h...	skin outermost organ becomes first protector h...	skin outermost organ becom first protector hum...	skin outermost organ become first protector hu...	skin outermost organ become first protector hu...	['skin, outermost, organ, become, first, protec...
4	2019	The writing of this scientific article aims to...	the writing of this scientific article aims to...	writing scientific article aims explain use fa...	writing scientific article aims explain use fa...	writing scientific article aims explain use fa...	write scientif articl aim explain use face det...	write scientific article aim explain use face ...	write scientific article aim explain use face ...	['write, scientific, article, aim, explain, use...

Gambar 6 *Tokenizing*
(Sumber: Google colab)

Tokenisasi biasanya melibatkan langkah-langkah seperti memisahkan kata-kata menggunakan spasi atau tanda baca sebagai pemisah. Namun, tokenisasi bisa lebih kompleks dalam beberapa bahasa atau dalam konteks tertentu.

2. Hasil *Topic modeling Hierarchical Dirichlet Process* (HDP)

Proses Topic modeling Hierarchical Dirichlet Process (HDP) dimulai dari hasil perolehan Data Preprocessing data. Setelah masuk pada Google colab peneliti membuat coding untuk mencari topic 1-10 terdiri dari kata apa saja dari masing-masing topic. Sedangkan secara manual perhitungan atau rumus HDP adalah sebagi berikut:

$$P(\theta) = DP(\alpha, H)$$

di mana θ adalah distribusi topik, DP(α,H) mengindikasikan bahwa θ adalah sampel dari DP dengan parameter konsentrasi α dan distribusi dasar (*base distribution*) H. Implementasi Python yang dijelaskan sebelumnya, untuk memperoleh model HDP menggunakan pustaka Gensim, rumus-rumus ini telah diimplementasikan secara internal oleh pustaka tersebut. Oleh karena

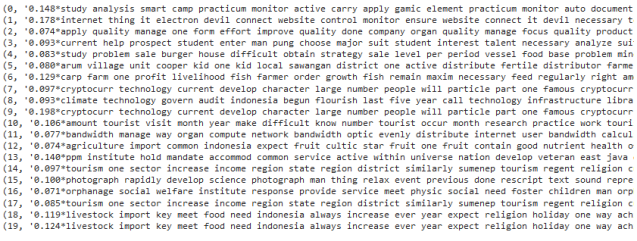
itu, ketika memanggil **HdpModel**, Gensim mengelola seluruh proses inferensi dan pembentukan model berdasarkan rumus-rumus di atas. Hal ini memudahkan pengguna untuk fokus pada penggunaan model dan hasilnya tanpa perlu secara eksplisit menentukan atau mengimplementasikan rumus-rumus tersebut secara manual.

Pemodelan tema menggunakan algoritma *Latent Dirichlet Allocation* (LDA) dengan jumlah topik sebanyak 20. LDA adalah salah satu metode yang umum digunakan dalam analisis tema pada dokumen teks[16] .Langkah pertama adalah menginisialisasi dan melatih model LDA dengan menggunakan dataset pelatihan (X_train) menggunakan fungsi fit dari objek model LDA.

Setelah melatih model, kita mendapatkan matriks topik-kata (topic_word_matrix), di mana setiap barisnya mewakili suatu topik dan setiap kolomnya mewakili kata dalam kosakata. Matriks ini menyimpan probabilitas distribusi kata-kata untuk setiap topik.

Selanjutnya, dilakukan ekstraksi kata-kata teratas untuk setiap topik, diurutkan berdasarkan probabilitasnya dalam matriks. Dengan mengakses fitur nama dari vektorisasi (vectorizer) yang digunakan pada pemrosesan teks sebelumnya, kita dapat mengidentifikasi kata-kata yang paling berkontribusi untuk setiap topik. Hasilnya ditampilkan dengan menggunakan perulangan, di mana setiap topik diikuti oleh daftar kata-kata teratasnya.

Proses ini membantu menerjemahkan hasil dari model LDA ke dalam bentuk yang dapat diinterpretasikan. Setiap topik yang dihasilkan oleh model LDA dapat diasosiasikan dengan sekumpulan kata-kata tertentu yang mencerminkan tema atau konsep tertentu. Dengan memahami kata-kata kunci ini, pengguna dapat memahami esensi dari masing-masing topik dan mengkategorikan dokumen-dokumen dalam dataset berdasarkan distribusi topiknya. Berdasarkan hasil *coding* diatas peneliti dapat melihat *Output* topic yang muncul, *Output* yang dihasilkan terdiri dari sepuluh *topic*, *Output* dari hasil *coding* tersebut ialah:



Gambar 7. Topik HDP
(Sumber: Google colab)

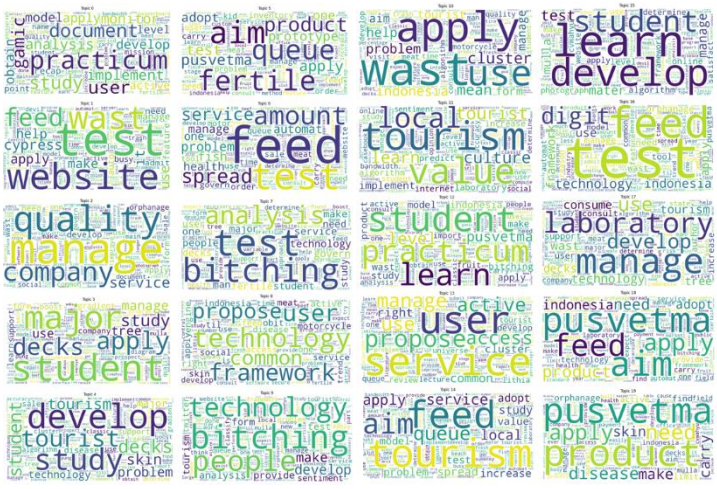
Dokumen yang telah dipetakan dalam 20 topik ini, selanjutnya dapat ditampilkan model yang berisikan peluang setiap kosa kata dalam masing-masing topik. Pembuatan model ini bertujuan untuk memudahkan pengguna apabila ada data dokumen baru yang ingin diketahui akan masuk klaster/topik mana. Penulisan model didasarkan pada, peluang untuk kata dalam setiap topik. Berdasarkan jumlah topik optimum, maka akan dibentuk model

peluang berjumlah 20 sebagai berikut.

$P(\varphi_1)$ = 0.148**study* + 0.066**photograph* + … + 0.031**submit*
$P(\varphi_2)$ = 0.178**internet* + 0.084**increase* + … + 0.029**technology*
$P(\varphi_3)$ = 0.074**apply* + 0.073**photograph* + … + 0.036**informant*
$P(\varphi_4)$ = 0. 093**current* + 0.070**manage* + … + 0.035**study*
$P(\varphi_5)$ = 0. 083**study* + 0.065**current* + … + 0.035**beef*
$P(\varphi_6)$ = 0. 080**arum* + 0.073**livestock* + … + 0.038**carp*
$P(\varphi_7)$ = 0.129**carp* + 0.079**manage* + … + 0.036**internet*
$P(\varphi_8)$ = 0. 097**cryptocurr* + 0.084**order* + … + 0.037**study*
$P(\varphi_9)$ = 0.093**climate* + 0.087**today* + … + 0.037**carp*
$P(\varphi_{10})$ = 0.198*cryptocurr + 0.067**manage* + … + 0.035**study*
$P(\varphi_{11})$ = 0.106**amount* + 0.082**increase* + … + 0.036**current*
$P(\varphi_{12})$ = 0.077**bandwidth* + 0.076*banyuwangi + … + 0.039**climate*
$P(\varphi_{13})$ = 0.074**agriculture* + 0.064**current* + … + 0.040**increase*
$P(\varphi_{14})$ = 0.140*ppm + 0.114**amount* + … + 0.035**study*
$P(\varphi_{15})$ = 0.097**tourism* + 0.087**carp* + … + 0.034**read*
$P(\varphi_{16})$ = 0.100**photograph* + 0.093**learn* + … + 0.034**informant*
$P(\varphi_{17})$ = 0.071**orphanage* + 0.067**current* + … + 0.034**beef*
$P(\varphi_{18})$ = 0.085**tourism* + 0.085**technology* + … + 0.038**bandwidth*
$P(\varphi_{19})$ = 0.119**livestock* + 0.073**carp* + … + 0.040**order*
$P(\varphi_{20})$ = 0. 124**livestock* + 0.073**orphanage* + … + 0.035*surabaya

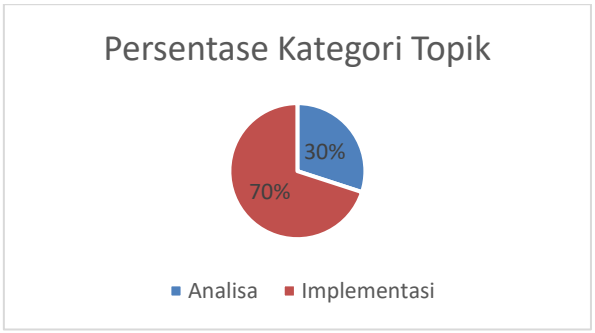
Berdasarkan pada persamaan diatas, maka dapat diketahui peluang kata pada setiap topik, dimana dinotasikan dengan φ yang merupakan distribusi kata ke-i pada topik ke-k. Jumlah topiknya adalah 20 yang merupakan jumlah optimum berdasarkan perhitungan *coherence*. Dari ke-20 model yang telah disajikan, maka dapat diketahui peluang masing-masing kata didalamnya. Sebagaimana yang telah disebutkan sebelumnya, bahwa tujuan dibuatnya model dalam metode ini untuk pemetaan jika digunakan dokumen baru. Jika ada suatu dokumen baru yang ingin diketahui dokumen tersebut masuk dalam topik mana, maka hal tersebut dapat dihitung. Langkah yang dilakukan yaitu dengan menghitung peluang setiap kata didokumen baru pada masing-masing topik. Selanjutnya dilakukan penjumlahan untuk seluruh kata didalam dokumen tersebut yang kemudian dipilih topik dengan 42 peluang terbesar. Sehinga dokumen baru akan masuk pada topik dengan peluang terbesar.

Berdasarkan hasil *Hierarchical Dirichlet Process* (HDP) peneliti dapat mengelompokkan dari masing-masing topik yang telah dihasilkan. Berikut merupakan wordCloud hasil pengelompokan berdasarkan penilaian peneliti.



Gambar 8. Word Cloud
(Sumber: Google colab)

Cara membaca gambar di atas yaitu dengan menentukan kata yang paling besar muncul pada setiap topik itu adalah besaran data apakah kata tersebut masuk kedalam kategori Analisa atau Implementasi, sehingga dari gambar di atas mayoritas kata yang sering muncul adalah data dengan kategori implementasi dengan data dari 20 topik sebagian besar terdapat 14 topik kedalam kategori implementasi dan sebagian kecil terdapat 6 topik masuk dalam kategori analisa.



Gambar 9. Persentase Perbandingan Jenis Topik
(Sumber: Data Statitistik Manual)

Dari Gambar 9, didapatkan persentase jenis topik yakni 30 persen dengan kategori analisa dan 70 persen dengan kategori implementasi. Hal ini dapat terbukti bahwa topik yang sering dibahas selama empat tahun terakhir menunjukkan bahwa lebih condong penelitian dengan kategori implementasi. Namun prosentase yang dihasilkan sangat berbeda jauh. Hal ini menunjukkan bahwa penelitian analisa masih bisa dijadikan pertimbangan karena tingkat prosentasenya masih 30 persen.

Hal ini menunjukkan bahwa penelitian dengan kategori analisa bisa dijadikan saran untuk Jurnal Ilmiah Teknologi Informasi dan Robotika dalam pemerataan penelitian. Sehingga dengan perbedaan hasil yang signifikan perlu adanya pemetaan yang dilakukan oleh pengelola jurnal Teknologi Informasi dan Robotika. Sehingga dengan adanya pemetaan dengan tujuan pemerataan penelitian dapat menjadikan rumah jurnal tersebut dapat di tingkatkan dari kualitas jurnalnya seperti peningkatan dalam tingaktan sinta atau scopus.

PENUTUP

Berdasarkan hasil penelitian dan pembahasan yang telah dijelaskan, peneliti dapat menarik kesimpulan sebagai berikut:

1. Penerapan pemodelan topik menggunakan metode Hierarchical Dirichlet Process (HDP) pada abstrak Jurnal Ilmiah Teknologi Informasi dan Robotika (JIFTI) dimulai dengan proses pengumpulan data. Sebanyak 42 abstrak jurnal dikumpulkan untuk digunakan dalam analisis. Data ini kemudian disiapkan melalui tahap pre-processing untuk memfasilitasi proses topic modelling. Pre-processing bertujuan untuk membersihkan dan mempersiapkan data agar lebih mudah diolah. Setelah itu, data yang telah dipersiapkan diinput ke dalam model topic modelling menggunakan metode HDP. Dalam analisis ini, fokus diberikan pada identifikasi 20 topik utama yang

ada dalam abstrak-abstrak tersebut. Proses ini melibatkan perhitungan kemunculan kata-kata dalam setiap abstrak, yang kemudian digunakan sebagai ukuran dalam model HDP.

2. Hasil topic modelling menggunakan metode HDP pada abstrak Jurnal Ilmiah Teknologi Informasi dan Robotika (JIFTI), dapat dilihat bahwa beberapa kata memiliki kemunculan yang cukup tinggi, seperti "practicum, test, manage, major, develop, aim, feed, bitching, technology, people, apply, tourism, student, user, feed, learn, digit, laboratory, pusvetma, product". Analisis terhadap kata-kata ini mengindikasikan bahwa mayoritas topik dalam abstrak JIFTI lebih berorientasi pada penelitian yang bersifat implementasi. Pernyataan ini didasarkan pada kata-kata kunci yang sering muncul dalam setiap topik yang dihasilkan dari analisis HDP, dan menunjukkan arah dominan dari penelitian yang dilakukan dalam jurnal tersebut.

Saran yang dapat diberikan. Pertama, untuk meningkatkan representasi topik yang dihasilkan oleh model, penting untuk memperluas dan memperkaya data yang digunakan dalam proses topic modelling. Mempertimbangkan teknik pre-processing yang lebih canggih untuk meningkatkan kualitas data yang digunakan dalam analisis. Selanjutnya, dalam pemodelan topik, dapat dipertimbangkan untuk menggunakan metode lain selain HDP, seperti Latent Dirichlet Allocation (LDA), untuk membandingkan hasil dan memastikan keakuratan dan ketepatan model.

Terakhir, agar hasil topic modelling dapat lebih bermanfaat, disarankan untuk melakukan interpretasi lebih lanjut terhadap setiap topik yang dihasilkan, dengan melibatkan ahli domain terkait untuk memahami implikasi dan relevansinya dalam konteks teknologi informasi dan robotika. Dengan demikian, proses topic modelling dapat memberikan wawasan yang lebih dalam dan berharga bagi para pembuat keputusan dalam bidang tersebut, sehingga dalam proses kategori lebih dipertajam dalam pemilihan kategori, tidak hanya menggunakan kategori analisa atau implementasi. Namun bisa juga pemilihan kategorinya lebih spesifik kepada metode penelitian yang digunakan pada Jurnal Ilmiah Teknologi Informasi dan Robotika (JIFTI).

DAFTAR PUSTAKA

[1] A. R. Rahim, *CARA PRAKTIS PENULISAN KARYA ILMIAH*. Yogyakarta: ZAHIR PUBLISHING, 2022.

[2] A. V. D. Sano, “Cara Kerja Data Mining - Seri Data Mining for Business Intelligence (3) | BINUS UNIVERSITY MALANG | Pilihan

Universitas Terbaik di Malang." Accessed: Apr. 30, 2024. [Online]. Available: https://binus.ac.id/malang/2019/01/cara-kerja-data-mining-seri-data-mining-for-business-intelligence-3/

[3] E. Turban, J. E. Aronson, and P. T. Liang, *Analisis topik data media sosial twitter menggunakan model topik Latent Dirichlet Allocation keke putri utami*. Connaught circus, 2005.

[4] A. Rokhim, Alimin, and M. L. Hakim, "Sistem Pendukung Keputusan Calon Penerima Dana Bantuan Siswa Miskin (Bsm) Menggunakan Metode Multi-Objective Optimazion on the Basis of Ratio Analysis," *Spirit*, vol. 14, no. 2, pp. 47–52, 2023, doi: 10.53567/spirit.v14i2.268.

[5] Y. Song and R. Wu, " The Impact of Financial Enterprises' Excessive Financialization Risk Assessment for Risk Control based on Data Mining and Machine Learning," *Comput Econ*, vol. 60, pp. 1245–1267, 2022.

[6] C. S. Hervilanda and R. Somya, "Perancangan Web Usage Mining Untuk Analisis Pola Pembelian Pelanggan di Online Shop ," Salatiga, Sep. 2019.

[7] C. C. Kelly, "DETECTING SUPPLIER PAYMENT ERRORS AND FRAUD USING A DATA WAREHOUSE AUDIT APPROACH," *EDPACS*, vol. 67, no. 3, pp. 1–20, 2023.

[8] F. R. Putra, "Data Mining dan Contoh Implementasi di Berbagai Bidang - Kompasiana.com." Accessed: Apr. 30, 2024. [Online]. Available: https://www.kompasiana.com/figorahput/5c927a740b531c34651a0062/datamining-dan-contoh-implementasi-di-berbagai-bidang.

[9] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2007.

[10] Y. Agusta, "K-Means – Penerapan, Permasalahan dan Metode Terkait ," *Jurnal Sistem dan Informatika*, vol. 3, pp. 47–60, Feb. 2007.

[11] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Boston: Pearson Education, 2006.

[12] J. Han, M. Kamber, and J. Pei, *Data Mining Concept and Techniques*, 3rd ed. Amsterdam: Elsevier, 2012.

[13] M. Mirnawati and F. Firman,

"Penerapan Teknik Clustering Dalam Mengembangkan Kemampuan Menulis Karangan Deskripsi Siswa Kelas IV MI Pesanten Datuk Sulaiman Palopo," *Jurnal Studi Guru dan Pembelajaran*, vol. 2, no. 2, pp. 165–177, May 2019, doi: 10.30605/JSGP.2.2.2019.1373.

[14] Zulhanif, Sudartianto, B. Tantular, and I. G. N. M. Jaya, " APLIKASI LATENT DIRICHLET ALLOCATION (LDA) PADA CLUSTERING DATA TEKS ," *Jurnal LOGIKA*, vol. 7, no. 1, pp. 46–51, 2017.

[15] D. M. Blei, A. Y. Ng, and J. B. Edu, "Latent Dirichlet Allocation Michael I. Jordan," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[16] M. E. Hoediansyah, "Pengelompokan Topik Skripsi Menggunakan Algoritma Agglomerative Hierarchical Clustering di Program Studi Sistem Informasi UPN Veteran Jawa Timur.," UPN Veteran, Surabaya, 2023.