Positif : Jurnal Sistem dan Teknologi Informasi
Volume 10, No 1, 2024, Page : 9-18

E-ISSN 2460-9552
P-ISSN 2620-3227

Article history
Received Nov 25, 2023
Accepted Feb 11, 2024
Published July 16, 2024

PEMANFAATAN METODE *TOPIC MODELLING HIERARCHICAL DIRICHLET PROCESS* DALAM MENGEVALUASI KUALITAS KONTEN *WEBSITE* BERDASARKAN ULASAN PENGGUNA

Sunu Jatmika[1)], Fransiska Sisilia Mukti[2)], Tria Aprilianto[3)], Naviza Yulia Al Zahwa[4)]
[1-4] Fakultas Teknologi dan Desain, Institut Teknologi dan Bisnis Asia Malang
email: sunu@asia.ac.id, ms.frans@asia.ac.id, tria@asia.ac.id, navizayulia@gmail.com

Abstract

The evaluation of website content is important to ensure that the presented content aligns with users' needs and preferences. This can be accomplished by analyzing user reviews regarding the website's content. This research leverages the Hierarchical Dirichlet Process (HDP) method to automatically identify primary topics from 32 users' reviews, resulting in three main recurring topics: 'good', 'bug', and 'update'. Using the OSEMN framework, the final evaluation indicates that the 'good' topic exhibits the highest cosine similarity value compared to other topics. This signifies that the positive aspects highlighted in users' reviews regarding the website's content dominate and possess significant similarities among the reviews. These findings offer crucial insights into comprehending user evaluations of website content, serving as a basis for more effective and targeted content improvements moving forward.

Keywords: *topic modelling, hierarchical dirichlet process, website content evaluation, user review analysis, cosine similarity*

Abstrak

Evaluasi konten *website* penting dilakukan untuk memastikan bahwa konten yang disajikan sesuai dengan kebutuhan dan preferensi pengguna. Hal ini dapat dilakukan melalui menganalisis hasil ulasan pengguna terhadap konten website. Penelitian ini memanfaatkan metode HDP dalam mengidentifikasi topik-topik utama secara otomatis dari ulasan 32 pengguna dan menghasilkan tiga topik utama yang paling sering muncul ‘bagus’, ‘bug’, ‘update’. Dengan menggunakan kerangka kerja OSEMN, evaluasi akhir menunjukkan bahwa topik 'bagus' memiliki nilai *cosine similarity* tertinggi dibandingkan dengan topik lainnya. Hal ini menandakan bahwa aspek positif dalam ulasan pengguna tentang kualitas konten website mendominasi dan memiliki kesamaan yang signifikan di antara ulasan-ulasan tersebut. Temuan ini memberikan wawasan yang penting dalam memahami evaluasi pengguna terhadap kualitas konten website dan dapat menjadi dasar untuk perbaikan konten yang lebih efektif dan terarah ke depannya.

Kata Kunci: pemodelan topik, hierarchical dirichlet process, evaluasi konten website, analisis ulasan pengguna, cosine similarity

1. PENDAHULUAN

Dalam ekosistem digital yang terus berkembang, *review* atau ulasan pengguna telah menjadi sumber daya kritis dalam membentuk persepsi dan keputusan konsumen. Informasi yang terkandung dalam ulasan ini mencakup beragam aspek, mulai dari kualitas produk hingga pengalaman pengguna. Oleh karena itu, memahami struktur dan konten dari ulasan-ulasan ini adalah esensial dalam mengidentifikasi preferensi dan kebutuhan pengguna.

Metode analisis teks tradisional sering kali menghadapi kendala dalam menangani kompleksitas dan dimensi yang berkembang dari dataset ulasan, terutama dalam konteks *review website* yang cenderung memiliki topik-topik yang beragam dan seringkali tidak terbatas. Oleh karena itu, ulasan secara *online* dianggap memberikan sumber data yang lebih kaya dibandingkan metode tradisional dalam hal memahami pengalaman holistik pelanggan dengan lebih baik [1].

Penggunaan alat analisis teks memiliki peran penting dalam membantu menyederhanakan serta mengotomatisasi proses pengambilan informasi dari *review customer*. Namun terdapat tantangan teknis yang dihadapi, antara lain akibat singkatan yang tidak standar, dialek lokal, ataupun kesalahan pengejaan [2].

Metode analisis teks, seperti *topic modelling*, adalah teknik penting dalam pengolahan dan pemahaman data teks. *Topic modelling* merupakan pendekatan statistik pada *text mining* yang digunakan untuk mengidentifikasi dan mengekstrak topik atau pola yang tersembunyi dalam kumpulan teks [3], [4]. Salah satu manfaat penggunaan metode *topic modelling* dalam *review website* adalah untuk mengidentifikasi sentimen atau perasaan umum pengguna yang terkait dengan produk atau layanan tertentu. Melalui metode ini, *web developer* dapat melihat apakah *review* yang disampaikan oleh pengguna cenderung positif, negatif, atau netral terhadap aspek-aspek tertentu.

Beberapa algoritma yang dikembangkan untuk *topic modelling* antara lain *latent dirichlet allocation* (LDA), *latent semantic analysis, hierarchical dirichlet process* (HDP), *correlated topic modeling,* dan *probabilistic latent semantic analysis* [5]. LDA adalah metode yang sudah terbukti dan banyak digunakan dalam pemodelan topik. Keuntungan dari LDA adalah kesederhanaannya dan kemampuannya untuk menghasilkan topik-topik yang dapat diinterpretasi dengan baik.

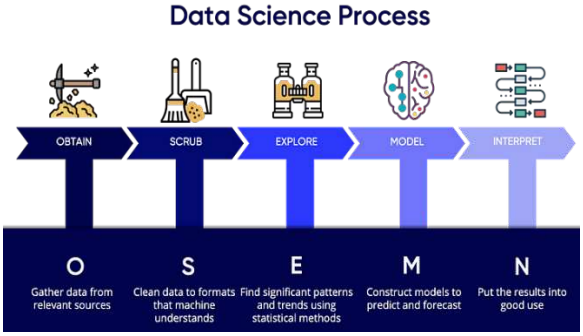
Di sisi lain, HDP adalah ekstensi dari LDA yang memungkinkan jumlah topik menjadi variabel dan dipelajari dari data. HDP secara otomatis menentukan jumlah topik yang sesuai dengan data, sehingga mengatasi kekurangan LDA di mana jumlah topik harus ditentukan sebelumnya. HDP juga dapat menangani situasi di mana dokumen memiliki campuran topik yang kompleks dan tidak terbatas [6].

Inilah alasan utama mengapa metode HDP mendapat perhatian yang signifikan. HDP membedakan dirinya dengan kemampuannya untuk menangani jumlah topik yang tidak terbatas, mengatasi tantangan yang umumnya dihadapi dalam metode konvensional. Dengan mengadopsi pendekatan hierarkis, HDP memungkinkan untuk memetakan ulasan-ulasan ke dalam struktur yang lebih dalam, mengidentifikasi topik utama serta sub-topik yang mungkin tidak terlihat dalam analisis konvensional [7].

Dengan demikian, penerapan HDP dalam analisis *review website* bukan hanya sekadar peningkatan metode analisis, melainkan sebuah langkah maju yang mampu menghadirkan pemahaman yang lebih mendalam dan kontekstual terkait preferensi serta kecenderungan pengguna dalam ranah digital yang semakin luas.

2. METODE PENELITIAN

Penelitian ini mengadopsi kerangka kerja OSEMN (*Obtain, Scrub, Explore, Model, dan iNterpret*), yang menyediakan panduan struktur untuk mengelola proyek *data science* dari awal hingga akhir. Pemilihan OSEMN sebagai metodologi penelitian merupakan langkah penting dalam memastikan bahwa penelitian ini dilakukan secara terstruktur dan komprehensif [8]. Gambar 1 menunjukkan penjelasan singkat mengenai OSEMN *framework.*



Gambar 1. OSEMN Framework [9]

E-ISSN 2460-9552
P-ISSN 2620-3227

A. OBTAIN

Pada tahap ini, fokus utama yang dilakukan adalah pengumpulan data dalam bentuk ulasan pengguna yang didapatkan dari hasil penyebaran kuesioner kepada sejumlah pengguna website, dengan jumlah sample sebanyak 32 pengguna. Hasil kuesioner ini disimpan dalam bentuk format *spreadsheet* yang menjadi data *input* untuk proses selanjutnya.

B. SCRUB

Proses *scrubbing* dalam siklus proyek *data science* dikenal sebagai *data preprocessing*, karena di dalamnya terkandung proses pembersihan data, khususnya untuk data dengan tipe teks yang cenderung tidak terstruktur dan terdapat banyak *noise*. Selain itu, pada proses *scrubbing* juga dilakukan konversi format data ke dalam satu standarisasi yang sama.

Tiga tahapan utama proses *scrubbing* dalam penelitian ini diuraikan sebagai berikut:

a) *Tokenization*: proses memecah teks atau kalimat menjadi unit-unit yang lebih kecil, yang disebut "token". Token bisa berupa kata, frasa, atau simbol, tergantung pada tingkat detail yang diinginkan. Tujuan dari tokenization adalah untuk mempersiapkan teks sehingga dapat diolah lebih lanjut dalam analisis.
b) *Stopwords:* kata-kata umum yang sering muncul dalam teks tetapi cenderung kurang informatif dalam konteks analisis teks. Menghapus stopwords dari teks dapat membantu mengurangi kebisingan dan meningkatkan kualitas analisis.
c) *Lemmatization:* proses dalam pengolahan bahasa alami yang mengubah kata-kata menjadi bentuk dasar atau kata dasar. Proses ini dilakukan untuk mengurangi kompleksitas dan dimensi dari dataset dan membantu dalam meningkatkan efisiensi analisis.

C. EXPLORE

Proses eksplorasi data untuk memberikan pemahaman secara mendalam terhadap data dalam mencari karakteristik dan tren. Salah satu teknik eksplorasi data yang dapat digunakan adalah TF-IDF (*Term Frequency-Inverse Document Frequency*). TF-IDF digunakan untuk mengekstrak kata-kata kunci yang paling penting dari dokumen. Kata-kata kunci ini kemudian dapat digunakan untuk mengelompokkan dokumen ke dalam topik-topik tertentu [10], [11].

D. MODEL

Setelah data dieksplorasi, langkah selanjutnya adalah membangun model untuk mengevaluasi kualitas konten website berdasarkan ulasan pengguna. Pada penelitian ini, metode yang digunakan adalah *Topic Modelling Hierarchical Dirichlet Process* (HDP). Metode HDP efektif digunakan untuk masalah yang melibatkan banyak kelompok data, dan metode ini merupakan ekstensi hirarkis dari *Dirichlet Process* (DP) [12]. Melalui metode HDP, ulasan pengguna *website* akan dikelompokkan ke dalam topik-topik tertentu untuk selanjutnya dilakukan evaluasi terhadap kualitas konten website berdasarkan topik-topik tersebut.

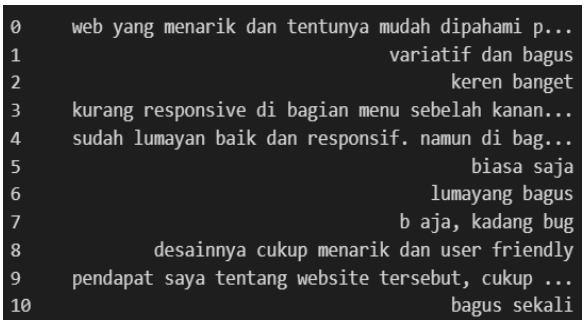
E. INTERPRET

Hasil dari pemodelan perlu divisualisasikan melalui tahapan *interpret.* Algoritma HDP membantu dalam proses ekstraksi istilah-intilah penting dalam hasil ulasan pengguna, dan proses visualisasi dibutuhkan untuk memberikan gambaran pemahaman yang lebih baik mengenai topik-topik individual dan hubungannya. Tujuan akhir dari penelitian ini adalah mengetahui masalah umum yang dialami pelanggan, sehingga dapat digunakan sebagai dasar untuk mengoptimalkan kualitas konten *website*.

3. HASIL DAN PEMBAHASAN

A. Data Pre-processing

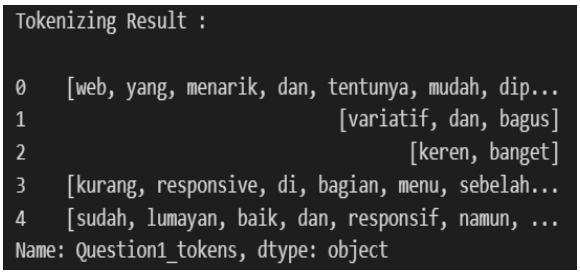
Kuesioner yang dirancang dalam penelitian mencakup pertanyaan yang relevan dengan evaluasi konten website, yang dapat menggali pemahaman dan persepsi pengguna terhadap konten yang diakses. Beberapa aspek yang dijadikan sebagai bahan pertanyaan pada kuesioner meliputi aspek kejelasan informasi, kerelevanan dengan topik, kegunaan, kepuasan pengguna, dan aspek lain yang relevan dengan evaluasi kualitas konten. Data yang dikumpulkan dari kuesioner tersebut akan menjadi dasar utama untuk melakukan analisis. Hasil kuesioner selanjutnya diolah menggunakan library *pandas python* sebagai data *input* metode HDP, sebagaimana yang terlihat pada Gambar 2.



Gambar 2. Tampilan Hasil Read Kuesioner

B. Tokenization

Hasil dari pembacaan data kuesioner akan membantu rangkaian kalimat yang tak terpisahkan. Proses *tokenizing* dilakukan untuk menguraikan setiap kalimat *review* pengguna ke dalam kata-kata, untuk memudahkan dalam proses perhitungan kata. Sebagai contoh *"web yang menarik…"* akan diubah menjadi ["web", "yang", "menarik", "…"]. Proses ini dilakukan pada seluruh data dan menghasilkan data baru seperti yang terlihat pada Gambar 3.



Gambar 3. Hasil Tokenization

C. Stopwords

Dalam memproses penghilangan kata-kata yang dianggap tidak penting, dibutuhkan *stoplist,* yaitu daftar kata umum yang mempunyai fungsi tapi tidak mempunyai arti, sebagaimana yang terlihat pada Gambar 4.

Gambar 4. Database Stopwords

Kata yang termasuk ke dalam stoplist akan dibuang dan tidak digunakan pada proses selanjutnya. Dalam penelitian ini, mekanisme pembersihan *stopwords* dilakukan dengan menggunakan metode *database* kamus *stopwords sastrawi*.

Kata yang dicetak miring dan berwarna merah merupakan kata-kata yang termasuk dalam *stoplist*, sehingga dari hasil *stopwords removal* akan menghapus kata tersebut, sebagaimana yang terlihat di Tabel 1 pada kolom di sebelah kanan.

D. Lemmatization

Proses penguraian kata-kata imbuhan menjadi kata dasar dengan tujuan untuk mendapatkan bentuk kata dasar yang benar. Proses ini dilakukan dengan menggunakan *SpaCy library python.* Hasil dari *lemmatization* dalam bentuk daftar kata dasar, sebagai contoh kalimat pada Tabel 1 *"UI dan UX nya mudah dipahami"* menjadi ['UI', 'UX', 'paham'].

E. Pembobotan Term dengan TF-IDF

Pada tahap ini merupakan tahap pembobotan yang dimana akan dilakukan pengubahan data yang berbentuk kata menjadi dalam bentuk numerik dengan menggunakan pembobotan TF-IDF. Pembobotan TF-IDF adalah gabungan dari metode *Term Frequency* (TF) dengan metode *Inverse Document Frequency* (IDF).

1) *Term Frequency* (TF)
   TF mengukur seberapa sering suatu kata muncul dalam suatu dokumen. Setiap kata pada dokumen akan diberi nilai 1 (jika muncul) dan 0 (jika tidak muncul) pada kolom TERM, dimulai dari term pada dokumen ke-1 (D1) hingga dokumen ke-32 . Sample hasil perhitungan TF pada dokumen kuesioner ditunjukkan melalui Gambar 5.

Positif : Jurnal Sistem dan Teknologi Informasi
Volume 10, No 1, 2024, Page : 9-18
E-ISSN 2460-9552
P-ISSN 2620-3227

	TF																															
TERM	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20	D21	D22	D23	D24	D25	D26	D27	D28	D29	D30	D31	D32
website	1	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0	1	0	0	0	0	1
tarik	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0
mudah	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
paham	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
guna	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
serta	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
userfriendly	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
variatif	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bagus	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	1	1	1	1	0	0	1	1	0	1
keren	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
banget	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kurang	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
responsive	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bagian	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
menu	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kanan	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
waktu	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
background	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
putih	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
tulisan	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ikut	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
lihat	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lumayan	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
baik	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2

Gambar 5. Sample Perhitungan TF pada Kolom ke-1

Untuk menghindari adanya anomaly data, maka dibutuhkan proses normalisasi terhadap perhitungan TF. Hal ini dapat dilakukan dengan menggunakan persamaan berikut.

$$TF(t,d) = \frac{jumlah\ kemunculan\ term\ t\ dalam\ dokumen\ d}{total\ jumlah\ term\ dalam\ dokumen\ d}$$

dimana, *t* adalah *term* (kata) yang dianalisis, dan *d* adalah dokumen yang sedang dipertimbangkan. Perhitungan ini memberikan nilai relatif antara 0 dan 1, yang menunjukkan frekuensi kemunculan kata dalam dokumen terhadap total kata dalam dokumen tersebut. Nilai TF yang lebih tinggi menunjukkan kata tersebut lebih penting atau relevan dalam dokumen. Gambar 6 menunjukkan hasil normalisasi dari data TF kolom ke-a.

TF NORMALISASI																															
D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20	D21	D22	D23	D24	D25	D26	D27	D28	D29	D30	D31	D32
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0,1
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,1

Gambar 4. Perhitungan Normalisasi untuk Kolom ke-1

2) *Inverse Document Frequency* (IDF)

IDF mengukur seberapa penting suatu kata dalam seluruh koleksi dokumen (korpus). Untuk dapat menghitung nilai IDF, terlebih dahulu harus dilakukan proses kalkukasi terhadap nilai DF, dengan menghitung kemunculan *term* atau kata yang muncul pada dokumen ke-1 hingga dokumen ke-32, dengan menggunakan persamaan berikut ini.

$$TF(t,d) = \frac{jumlah\ kemunculan\ term\ t\ dalam\ dokumen\ d}{dokumen\ ke-n}$$

Hasil perhitungan DF dalam penelitian ini ditunjukkan melalui Gambar 7. Hanya data dengan nilai *term* > 0 yang akan dimunculkan.

DOCUMENT FREQUENCY (DF)																														
8	5	3	2	1	2	2	1	12	1	1	4	4	2	3	2	1	2	2	1	3	3	3	7	1	5	2	2	1	1	1	2
2	1	1	2	1	1	1	1	1	1	1	1	2	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	3	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
1	1	1	1	1	1	3	1																								

Gambar 5. Hasil Perhitungan DF

Setelah mengetahui nilai DF, maka perhitungan IDF dapat dengan mudah dilakukan dengan menggunakan persamaan berikut ini.

$$IDF_t = \log(\frac{N}{DF_t})$$

dimana, *N* merupakan jumlah keseluruhan dokumen. Perhitungan ini memberikan nilai yang lebih tinggi untuk kata-kata yang jarang muncul dalam seluruh koleksi dokumen, menunjukkan kata-kata yang lebih unik dan mungkin lebih penting. Hasil perhitungan IDF ditunjukkan melalui Gambar 8.

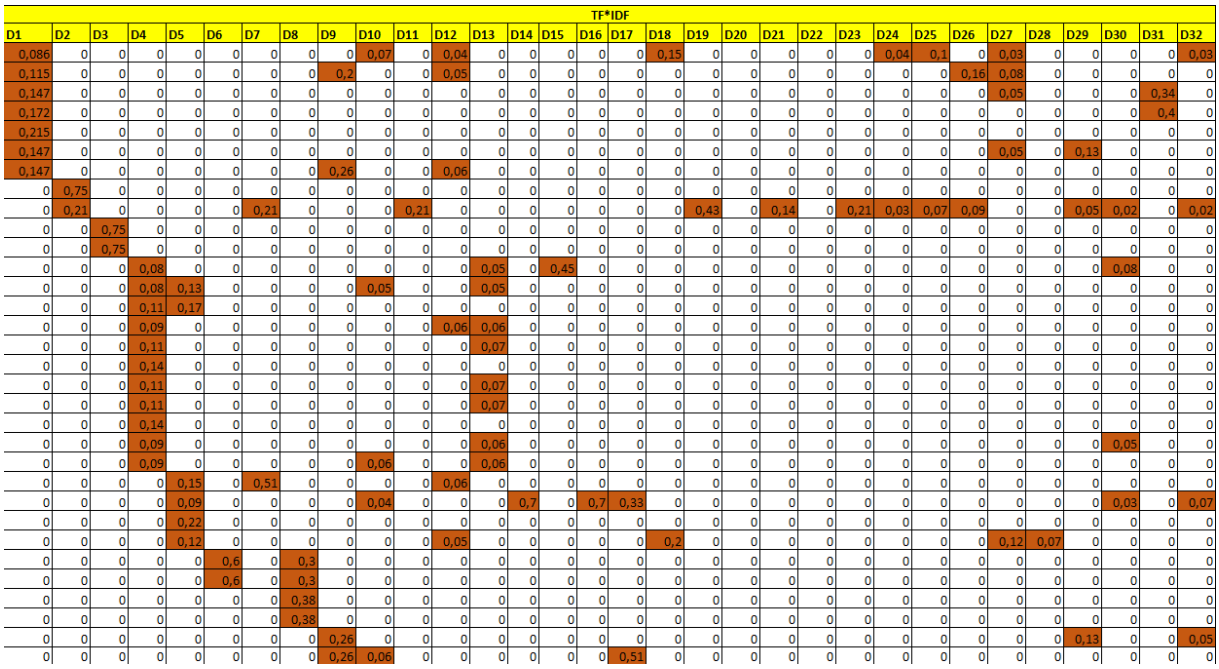
INVERSE DOCUMENT FREQUENCY (IDF)													
0,6	0,81	1,03	1,2	1,51	1,03	1,03	1,51	0,43	1,51	1,51	0,9		
0,9	1,2	1,03	1,2	1,51	1,2	1,2	1,2	1,51	1,03	1,03	1,03	0,66	1,51
0,81	1,2	1,2	1,51	1,03	1,03	1,2	1,2	1,03	0,73	0,9	1,51	1,51	0,9
0,9	1,2	1,2	1,51	1,51	1,51	1,2	1,2	1,51	1,51	1,2	1,51	1,2	1,51
1,51	1,51	1,51	1,51	1,51	1,51	1,51	1,51	1,2	1,51	1,2	1,51	1,51	1,2
1,51	1,51	1,51	1,51	1,51	1,51	1,51	1,51	1,51	1,51	1,51	0,9	1,51	1,51
1,51	1,51	1,51	1,03	1,2	1,51	1,51	1,51	1,51	1,51	1,51	1,51	1,2	1,51
1,51	1,51	1,51	1,51	1,51	1,51	1,51	1,51	1,51	1,51	1,51	1,51	1,2	1,51
1,51	1,51	1,51	1,51	1,03	1,51								

Gambar 6. Hasil Perhitungan IDF

3) TF-IDF

Setelah menghitung TF dan IDF, nilai TF-IDF dari suatu *term* dalam suatu dokumen dapat diperoleh dengan mengalikan nilai TF dengan nilai IDF. Perhitungan ini memberikan bobot pada kata-kata yang sering muncul dalam dokumen tersebut (TF yang tinggi), tetapi juga mempertimbangkan seberapa unik kata tersebut dalam seluruh koleksi dokumen (IDF yang tinggi). Hasil perhitungan TF-IDF dari keseluruhan dokumen ditunjukkan melalui Gambar 9.

E-ISSN 2460-9552
P-ISSN 2620-3227



TF*IDF																														
D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20	D21	D22	D23	D24	D25	D26	D27	D28	D29	D30	D31	D32
0,086	0	0	0	0	0	0	0	0	0,07	0	0,04	0	0	0	0	0	0,15	0	0	0	0	0	0,04	0,1	0	0,03	0	0	0	0	0,03
0,115	0	0	0	0	0	0	0	0,2	0	0	0,05	0	0	0	0	0	0	0	0	0	0	0	0	0	0,16	0,08	0	0	0	0	0
0,147	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,05	0	0	0	0,34	0
0,172	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,4	0
0,215	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0,147	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,05	0	0,13	0	0	0
0,147	0	0	0	0	0	0	0	0,26	0	0	0,06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0,75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0,21	0	0	0	0	0,21	0	0	0	0,21	0	0	0	0	0	0	0	0,43	0	0,14	0	0,21	0,03	0,07	0,09	0	0	0,05	0,02	0	0,02
0	0	0,75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0,75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0,08	0	0	0	0	0	0	0	0	0,05	0	0,45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,08	0	0
0	0	0	0,08	0,13	0	0	0	0	0,05	0	0	0,05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0,11	0,17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0,09	0	0	0	0	0	0	0	0,06	0,06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0,11	0	0	0	0	0	0	0	0	0,07	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0,14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0,11	0	0	0	0	0	0	0	0	0,07	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0,11	0	0	0	0	0	0	0	0	0,07	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0,14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0,09	0	0	0	0	0	0	0	0	0,06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,05	0	0
0	0	0	0,09	0	0	0	0	0	0,06	0	0	0,06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0,15	0	0,51	0	0	0	0	0,06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0,09	0	0	0	0	0,04	0	0	0	0,7	0	0,7	0,33	0	0	0	0	0	0	0	0	0	0	0	0	0,03	0	0,07
0	0	0	0	0,22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0,12	0	0	0	0	0	0	0,05	0	0	0	0	0	0,2	0	0	0	0	0	0	0	0	0,12	0,07	0	0	0	0
0	0	0	0	0	0,6	0	0,3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0,6	0	0,3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0,38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0,38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0,26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,13	0	0	0,05
0	0	0	0	0	0	0	0	0,26	0,06	0	0	0	0	0	0	0,51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Gambar 7. Sample Pehitungan TF-IDF pada Kolom ke-1

F. Cosine Similarity (CS)

Setelah mendapatkan nilai pembobotan pada masing-masing *term,* selanjutnya dilakukan perangkingan dokumen untuk menghitung nilai *cosine similarity,* yaitu metode yang menghitung tingkat kemiripan antara dua objek atau lebih.

Dalam HDP atau model-topic-based lainnya, setiap dokumen direpresentasikan sebagai vektor distribusi topik, di mana setiap koordinat vektor menunjukkan seberapa banyak topik tertentu memengaruhi dokumen tersebut.

Berdasarkan proses pembobotan TF-IDF pada penjelasan sebelumnya, didapatkan 3 topik atau *keywords* yang dihasilkan dalam penelitian ini, yaitu [*'bagus', 'bug', 'update'*]. Dari ketiga topik ini, selanjutnya akan dilakukan perhitungan *keywords* atau kata kunci dari dokumen yang tersedia, dengan cara mencari nilai TF-IDF untuk masing-masing *keywords,* untuk kemudian dicari nilai vektornya dengan menggunakan persamaan berikut ini

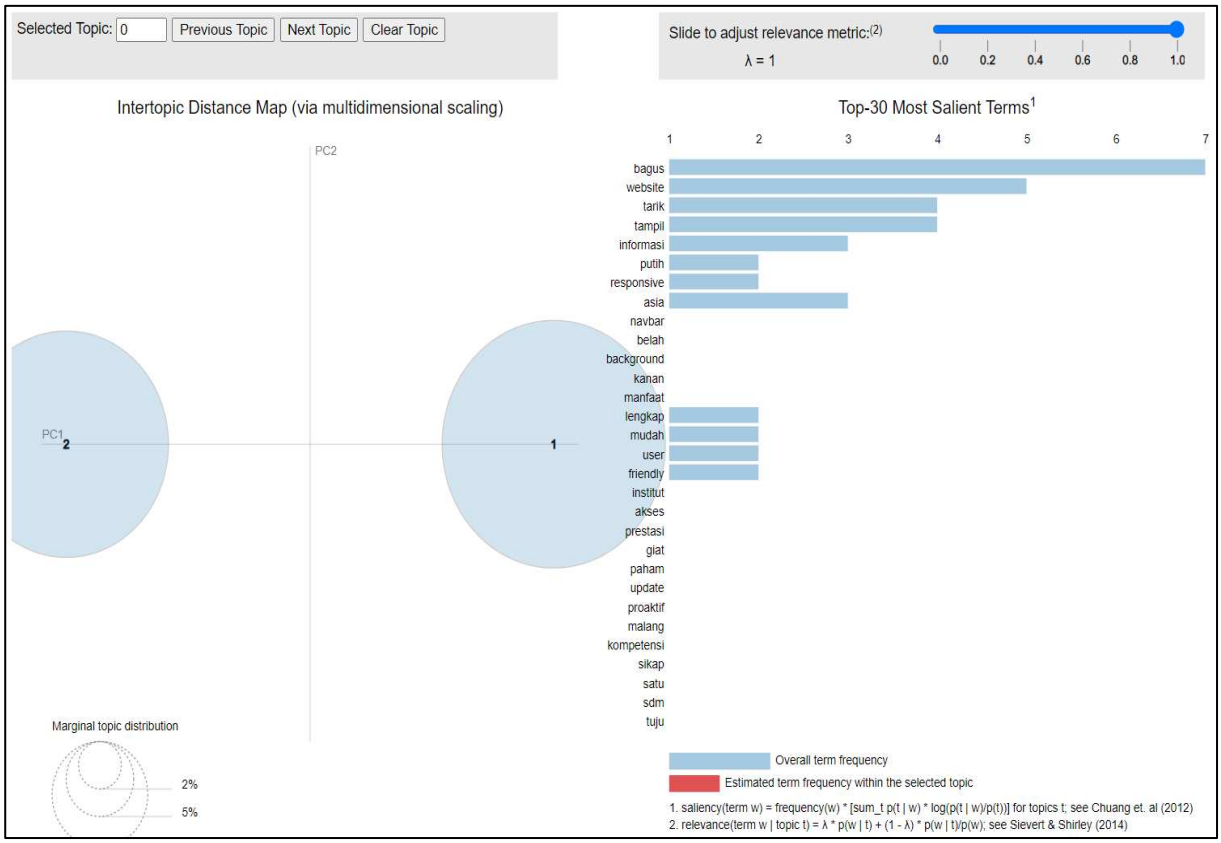
$$KK = \sqrt{term1^2 + term2^2 + term3^2}$$

Berdasarkan hasil perhitungan TF-IDF pada tahapan sebelumnya, didapatkan hasil *value* dari *keywords* ['bagus'] sebesar 0,43; *value keywords* ['bug'] sebesar 1,51; dan *value keywords* ['update'] sebesar 0,9. Maka, dari persamaan di atas, didapatkan nilai vector *keywords (KK)* sebesar 0,584 yang nantinya akan digunakan sebagai vector distribusi topik pada perhitungan CS.

Cosine similarity digunakan dalam penelitian ini sebagai metrik yang efektif untuk mengukur kesamaan antara vektor-vektor yang mewakili dokumen-dokumen tersebut. Perhitungan *cosine similarity* antara dua vektor distribusi topik (misalnya, dua dokumen) dilakukan dengan menggunakan persamaan berikut ini:

$$CS\ (A, B) = \frac{A\ .\ B}{\|A\| \times \|B\|}$$

dimana *A* menunjukkan nilai vector *KK* dan *B* menunjukkan nilai vector dari setiap dokumen. Hasil dari perhitungan tingkat kemiripan setiap dokumen dengan *KK* ditunjukkan melalui Gambar 10.



Gambar 8. Visualisasi Akhir HasilPemodelan Topik Review Website

G. INTERPRET

Langkah terakhir dalam pemodelan topik menggunakan metode HDP dalam penelitian ini ditunjukkan dalam bentuk visualisasi hasil *review* pengguna. Terdapat tiga topik utama yang dihasilkan berdasarkan kuesioner pengguna, dan *term* [‘bagus’] menunjukkan tingkat probabilitas tertinggi sebagaimana yang terlihat pada Gambar 11.

D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
0	0,2	0	0	0	0	0,282	1,427	0	0
D11	**D12**	**D13**	**D14**	**D15**	**D16**	**D17**	**D18**	**D19**	**D20**
0,2	0	0	0	0	0	0	0	0,736	0

D21	D22	D23	D24	D25	D26	D27	D28	D29	D30	D31	D32
0,158	0	0,282	0,064	0,126	0,125	0	0	0,087	0,05	0	0,057

Gambar 9. Perhitungan CS untuk Seluruh Dokumen

Hasil pemodelan topik menggunakan HDP perlu untuk divisualisasikan dengan tujuan memberikan pemahaman perspektif pengguna dengan lebih baik, mengidentifikasi area-area yang perlu diperbaiki, dan membuat keputusan yang lebih terarah untuk meningkatkan kualitas konten dan pengalaman pengguna untuk pengelola *website*.

Temuan ini memberikan pemahaman yang mendalam bahwa kebanyakan pengguna memberikan ulasan positif tentang kualitas konten website, yang dapat menjadi fokus utama untuk dipertahankan atau ditingkatkan. Oleh karena itu, pemilik website dapat menggunakan temuan ini sebagai dasar untuk mengembangkan strategi yang bertujuan untuk memperkuat dan mempertahankan aspek-aspek yang dinilai 'bagus' oleh pengguna, sambil terus memperbaiki area-area lain yang mungkin memerlukan perhatian lebih lanjut seperti aspek 'bug' atau 'update'.

E-ISSN 2460-9552
P-ISSN 2620-3227

4. PENUTUP

Kesimpulan

Evaluasi kualitas konten *website* penting untuk dilakukan demi meningkatkan kepuasa pengguna terhadap *website,* baik dari segi konten maupun secara visual. Banyaknya hasil ulasan yang diberikan maupun variasi data yang dihasilkan menbuat metode HDP menjadi salah satu *tools* yang cukup efisien dalam melakukan proses evaluasi dengan cepat dan tepat.

Dalam penelitian ini, melibatkan 32 pengguna dalam pengisian kuesioner memberikan wawasan yang berharga terhadap aspek-aspek kunci yang dibahas dalam ulasan mereka. Dari analisis HDP, teridentifikasi tiga topik utama yang mencakup aspek 'bagus', 'bug', dan 'update'. Namun, hasil kajian cosine similarity menunjukkan bahwa topik 'bagus' memiliki nilai cosine similarity tertinggi dibandingkan dengan topik lainnya. Hal ini menandakan bahwa aspek-aspek positif yang diungkapkan oleh pengguna dalam ulasan mereka mengenai kualitas konten website lebih dominan dan memiliki kesamaan yang lebih tinggi di antara ulasan-ulasan tersebut.

Dengan demikian, pemanfaatan metode HDP dalam evaluasi ulasan pengguna dapat memberikan wawasan yang berharga bagi pengambil keputusan dalam meningkatkan kualitas konten website secara keseluruhan.

Saran

Sebagai kelanjutan penelitian, dibutuhkan penambahan kedalaman analisis dengan menggabungkan metode analisis sentimen pada setiap topik yang diidentifikasi untuk memperkaya hasil penelitian dengan menyediakan gambaran yang lebih rinci tentang aspek mana yang dianggap positif, negatif, atau netral oleh pengguna. Selanjutnya, dibutuhkan cakupan data yang lebih luas dengan menambahkan jumlah ulasan pengguna.

5. UCAPAN TERIMA KASIH

Ucapan terima kasih ditujukan kepada Institut Asia Malang, melalui Lembaga Penelitian, Pengembangan dan Pengabdian kepada Masyarakat (LPPM) yang telah memberi dukungan secara finansial untuk terselesaikannya penelitian ini, dalam program hibah penelitian internal dengan nomor kontrak 0098/B.1/LP2M/ITB-ASIA/III/2023.

5. REFERENSI

[1] I. Sutherland, Y. Sim, S. K. Lee, J. Byun, and K. Kiatkawsin, “Topic Modeling of Online Accommodation Reviews via Latent Dirichlet Allocation,” *Sustainability*, vol. 12, no. 5, p. 1821, Feb. 2020, doi: 10.3390/su12051821.

[2] A. Djuraidah, B. Sartono, and Y. Putranto, “Topic Modelling and Hotel Rating Prediction based on Customer Review in Indonesia,” *International Journal of Management and Decision Making*, vol. 20, no. 1, p. 1, 2021, doi: 10.1504/IJMDM.2021.10036033.

[3] M. L. C. Chilmi, “Latent Dirichlet Allocation (LDA) untuk Mengetahui Topik Pembicaraan Warganet Twitter tentang Omnibus Law,” Jakarta, 2021.

[4] H. Jelodar *et al.*, “Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey.”

[5] I. Vayansky and S. A. P. Kumar, “A review of topic modeling methods,” *Inf Syst*, vol. 94, p. 101582, Dec. 2020, doi: 10.1016/j.is.2020.101582.

[6] K. Jeong and Y. Kim, “Dynamic hierarchical Dirichlet processes topic model using the power prior approach,” *J Korean Stat Soc*, vol. 50, no. 3, pp. 860–873, Sep. 2021, doi: 10.1007/s42952-021-00129-1.

[7] H. Zhang, S. Huating, and X. Wu, “Topic model for graph mining based on hierarchical Dirichlet process,” *Stat Theory Relat Fields*, vol. 4, no. 1, pp. 66–77, Jan. 2020, doi: 10.1080/24754269.2019.1593098.

[8] H. Mason and C. Wiggins, “A Taxonomy of Data Science.” Accessed: Nov. 07, 2023. [Online]. Available: http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

[9] C. H. Lau, “5 Steps of a Data Science Project Lifecycle,” Towards Data Science. Accessed: Nov. 07, 2023. [Online]. Available: https://towardsdatascience.com/5-steps-of-a-

data-science-project-lifecycle-26c50372b492

[10] M. Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure,” Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.05794

[11] L. George and P. Sumathy, “An integrated clustering and BERT framework for improved topic modeling,” *International Journal of Information Technology*, vol. 15, no. 4, pp. 2187–2195, Apr. 2023, doi: 10.1007/s41870-023-01268-w.

[12] H. Zhang, S. Huating, and X. Wu, “Topic model for graph mining based on hierarchical Dirichlet process,” *Stat Theory Relat Fields*, vol. 4, no. 1, pp. 66–77, Jan. 2020, doi: 10.1080/24754269.2019.1593098.