Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 4 Iss. 2 April 2024, pp: 481-489
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Topic Modeling on Online News Media Using Latent Diriclet Allocation (Case Study Somethinc Brand)

Pemodelan Topik pada Media Berita *Online* Menggunakan *Latent Dirichlet Allocation* (Studi Kasus Merek Somethinc)

Evi Puspita[1*], Diqy Fakhrun Shiddieq[2], Fikri Fahru Roji[3]

[1,2,3] Program Studi Bisnis Digital, Fakultas Ekonomi, Universitas Garut, Indonesia

E-Mail: [1]24025120033@fekon.uniga.ac.id,
[2]diqy@uniga.ac.id, [3]fikri@uniga.ac.id

Received Nov 21th 2023; Revised Jan 21th 2024; Accepted Feb 19th 2024
Corresponding Author: Evi Puspita

Abstract

Somethinc is one of the local cosmetic brands in Indonesia that actively uses media, such as online news, to convey the latest information about the brand. Among the numerous online news articles about the Somethinc brand, often the discussed topics and trends often do not comprehensively depict information. To analyze the most discussed topics in online news about the Somethinc brands, researchers used the topic modeling method, namely Latent Dirichlet Allocation, which is considered superior for generating structured topics. This research utilized coherence values to analyze and evaluate the optimal number of topics, followed by a human judgment approach to interpret the topics. The analysis results were then interactively visualized using pyLDAvis to understand the distribution of words for each topic. Based on the research findings, the optimal number of topics was found to be 6, with a coherence value of 0.404. These six topics were subsequently interpreted based on the human judgment approach, resulting in topics including skincare products for acne-prone skin, awards for best beauty brands, product collaborations, skincare and beauty products, product marketing campaigns and local brands with beauty care products. It can be concluded that the 6 topics generated relevant discussions about the Somethinc brand.

Keyword: Latent Dirichlet Allocation, Online News, Media, Topic Modeling

Abstrak

Somethinc merupakan salah satu merek kosmetik lokal di Indonesia yang aktif memanfaatkan media, seperti berita online untuk menyampaikan informasi terkini seputar merek. Dari banyaknya berita online mengenai merek Somethinc, sering kali topik dan tren yang sedang dibahas tidak menggambarkan informasi secara keseluruhan. Untuk menganalisis topik yang paling sering dibahas dalam berita online mengenai merek Somethinc, peneliti menggunakan metode topic modeling, yaitu Latent Dirichlet Allocation, yang dinilai lebih unggul dalam menghasilkan topik secara terstruktur. Penelitian ini memanfaatkan nilai *coherence* untuk menganalisis dan mengevaluasi jumlah topik terbaik, selanjutnya pendekatan *human judgement* digunakan untuk menginterpretasikan topik. Hasil analisis kemudian divisualisasikan secara interaktif menggunakan pyLDAvis, untuk mengetahui persebaran kata dari setiap topik. Berdasarkan hasil penelitian, jumlah topik terbaik terdapat pada topik 6 dengan nilai *coherence* sebesar 0.404. Keenam topik tersebut diinterpretasikan berdasarkan pendekatan *human judgement*, menghasilkan topik-topik meliputi produk *skincare* untuk kulit berjerawat, penghargaan *brand* kecantikan terbaik, kolaborasi produk, produk perawatan kulit dan kecantikan, kampanye pemasaran produk, dan *brand* lokal dengan produk perawatan kecantikan. Dapat disimpulkan bahwa jumlah topik 6 menghasilkan topik-topik yang relevan mengenai merek Somethinc.

Kata Kunci: Berita Online, Latent Dirichlet Allocation, Media, Pemodelan Topik

1. PENDAHULUAN

Kosmetik merupakan salah satu industri yang berperan sebagai penggerak utama perekonomian negara oleh Kementerian Perindustrian Indonesia [1]. Sejalan dengan pertumbuhan jumlah pelaku usaha industri kosmetik lokal yang meningkat secara signifikan, awalnya berjumlah 819 pada tahun 2021 menjadi 913 atau sebesar 20,5% pada tahun 2022 [2]. Pertumbuhan industri kosmetik juga didorong oleh

pertumbuhan populasi masyarakat serta kesadaran akan kesejahteraan diri [3]. Meskipun prospek bisnis kosmetik di Indonesia dianggap cukup besar, hal ini juga bisa menjadi ancaman berupa meningkatnya persaingan antara industri kosmetik lokal [4].

Somethinc merupakan salah satu merek kosmetik lokal yang didirikan pada tahun 2019 oleh PT. Royal Pesona Indonesia. Adapun beberapa penghargaan yang berhasil diraih oleh Somethinc yaitu *Best Cushion* (Female Daily Beauty Awards 2023), *Best Powder* (Tokopedia Beauty Awards 2023), *Best Facial Wash, Best Moisturizer, Best Eyebrow* dan *Best Eyeliner Product* (BeautyHaul Awards 2023) dan *Best Moisturizer* (Popbela Beauty Awards 2023). Somethinc telah diakui oleh Lembaga Prestasi Indonesia Dunia (LEPRID) pada tahun 2021 sebagai merek dengan pendaftaran tercepat yakni 1 bulan 8 hari dan mereknya telah terdaftar di 123 negara. Prestasi ini signifikan bagi Somethinc untuk dijadikan objek penelitian, karena ini adalah merek yang relatif baru didirikan dibandingkan dengan banyak pesaingnya.

Seiring perkembangan teknologi, Somethinc secara aktif memanfaatkan media *online* sebagai sarana untuk menyebarkan informasi mengenai produk, tren terbaru, perkembangan merek, dan konten edukatif seputar kecantikan. Saat ini, berita *online* merupakan salah satu media yang banyak dimanfaatkan oleh masyarakat untuk mengetahui perkembangan informasi terkini [5], [6]. Hal ini berarti, media berita *online* sering digunakan sebagai sumber referensi utama bagi masyarakat [7]. Perusahaan kosmetik, khususnya Somethinc membutuhkan peran dari berita *online*, untuk mengetahui penyampaian berupa interpretasi dari pemberitaan media [8] terhadap mereknya. Setiap portal berita memiliki pendekatan tersendiri dalam menyajikan informasi. Namun, dalam konteks berita *online* mengenai merek Somethinc, sering kali topik yang sedang dibahas tidak menggambarkan informasi secara keseluruhan. Informasi tersebut memerlukan analisis lebih mendalam. Di samping itu, media berita *online* memiliki banyak informasi sehingga peneliti membutuhkan metode yang cepat dan efisien untuk menganalisis topik yang paling sering dibahas.

Topic modeling merupakan salah satu teknik *unsupervised machine learning* untuk menemukan tema tersembunyi dari kumpulan dokumen teks besar guna mengelompokkan tema-tema tersebut menjadi satu topik [9], [10]. Pertama kali diperkenalkan oleh David Blei, *Latent Dirichlet Allocation* (LDA) merupakan salah satu pemodelan topik yang paling populer, LDA digunakan untuk menampilkan sebuah topik menggunakan probabilitas dari setiap kata sehingga dapat membantu menggambarkan dokumen-dokumen menjadi lebih terstruktur [11]. Penerapan *topic modeling* menggunakan LDA dianggap lebih unggul dalam menghasilkan topik bermakna logis [12] kemampuan interpretasi, dan performansi prediktif [13].

Penelitian terkait LDA telah dilakukan oleh peneliti sebelumnya untuk menganalisis topik pandemi Covid-19 di Indonesia menggunakan data yang bersumber dari Wikipedia [14]. Penelitian ini melakukan klasterisasi topik dengan menghitung probabilitas kata terhadap topik dari setiap iterasinya. Hasilnya menunjukkan bahwa jumlah *cluster* terbaik terdapat pada topik 3 dan menghasilkan topik relevan dengan kesehatan. Selanjutnya, metode LDA digunakan untuk mengetahui topik utama dari setiap cuitan akun @jokowi selaku pejabat negara pada media sosial Twitter [15]. Hasil analisis topik dievaluasi berdasarkan nilai *perplexity* dan nilai *coherence*. Ditemukan bahwa jumlah topik terbaik terdapat pada topik 7.

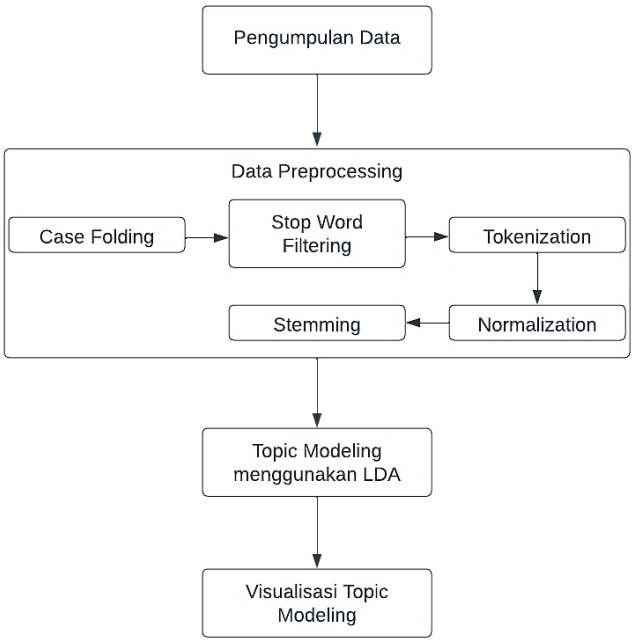
Dalam penelitian lain, dilakukan analisis untuk mengetahui topik yang dominan dibahas melalui tagar covidindonesia pada media sosial Instagram menggunakan metode LDA [16]. Penelitian tersebut melakukan evaluasi nilai *perplexity* dan nilai *coherence* untuk menemukan jumlah topik terbaik. Jumlah topik terbaik terdapat pada topik 3 sehingga menghasilkan topik seperti covidindonesia, covid_19, pandemic di Indonesia, dan pembahasan mutasi virus covid-19. Penelitian lainnya [17] menggunakan LDA untuk menguji apakah LDA dapat diterapkan pada tiket dukungan. Analisis menggunakan 5 parameter secara acak yaitu *number of topic*, nilai *alpha*, nilai beta, iterasi, dan *random_seed* untuk mendapatkan nilai *coherence* tertinggi. Kemudian dievaluasi menggunakan pendekatan penilaian manusia. Dalam penelitian ini, nilai coherence dan pendekatan penilaian manusia berpengaruh dalam menginterpretasikan topik, dengan demikian LDA terbukti dapat diterapkan pada tiket dukungan.

Sebagai perbedaan dengan penelitian sebelumnya, penelitian ini melakukan pengambilan data pada berita online khususnya berkaitan dengan merek kosmetik yaitu Somethinc. Evaluasi penelitian ini menggunakan nilai *coherence* dan pendekatan *human judgement* untuk menganalisis jumlah topik terbaik. Keterbaruan dari penelitian ini, terletak pada hasil dari pemodelan topik yang akan divisualisasikan secara interaktif menggunakan pyLDAvis untuk memberikan kemudahan mengetahui distribusi kata dari setiap topik. Tujuan dari penelitian ini adalah untuk menganalisis topik yang paling sering dibahas dalam media berita *online* mengenai merek Somethinc. Penelitian ini memberikan manfaat untuk menggali wawasan lebih mendalam terkait merek kosmetik khususnya Somethinc. Hasil dari penelitian ini dapat dijadikan sebagai sumber informasi bagi perusahaan untuk meningkatkan kualitasnya di masa mendatang. Disamping itu, masyarakat juga mendapatkan informasi yang relevan terkait merek kosmetik tersebut.

2. METODOLOGI PENELITIAN

Tahapan penelitian ini ditunjukkan pada Gambar 1. Tahapan pertama yaitu pengumpulan data mengenai merek Somethinc pada media berita *online*. Kemudian dilanjutkan dengan tahapan data *preprocessing* untuk membersihkan kata-kata yang tidak diperlukan. Tahap selanjutnya melakukan

pemodelan topik menggunakan *Latent Dirichlet Allocation* (LDA). Tahap terakhir pembuatan visualisasi dari hasil pemodelan topik menggunakan pyLDAvis.



Gambar 1. Alur Penelitian

2.1. Pengumpulan Data

Penelitian ini menggunakan data dari media berita *online*. Portal berita yang digunakan sebagai sumber pada penelitian ini dilihat berdasarkan reputasi media dan kredibilitas media dalam menyampaikan berita. Media berita *online* yang penelitian ini gunakan sebagai sumber data diantaranya yaitu fimela.com, kompas.com, kumparan.com, marketeers.com, antaranews.co, detik.com, dan tempo.co. Pengumpulan data dilakukan dengan menggunakan *search engine hacks* pada Google. Penelitian ini menggunakan operator *site* untuk membatasi hasil pencarian dan fokus pada satu situs web atau domain tertentu. Di samping itu, penelitian ini memanfaatkan operator *intitle* untuk mempersempit hasil pencarian pada halaman web dengan menggunakan kata kunci yang telah ditentukan oleh penelitian ini yaitu Somethinc. Penelitian ini juga memanfaatkan *tools* pada *search engine hacks* yaitu filter rentang waktu, sehingga berita mengenai merek Somethinc menjadi lebih spesifik. Setelah itu, penelitian ini menyeleksi dan menyalin isi berita berupa teks secara manual. Data yang telah dikumpulkan kemudian disimpan ke dalam file dalam format xlsx.

2.2. *Preprocessing* Data

Dalam penelitian ini, penelitian ini melakukan *preprocessing* data untuk memahami dan mengidentifikasi masalah terkait data, sehingga data dapat lebih diandalkan untuk kemudian digunakan [18]. Pada tahapan ini penelitian ini memanfaatkan bahasa pemrograman menggunakan *software* berbasis *Python*. Adapun tahapan *preprocessing* data diantaranya yaitu:

1. *Case folding* merupakan tahap awal untuk mengubah data ke format standar dengan mengkonversi huruf menjadi huruf kecil (*lowercase*). Selanjutnya dilakukan *cleaning data* untuk menghilangkan karakter seperti tanda baca, menghapus data duplikat, dan data yang tidak konsisten.
2. *Normalization* merupakan tahapan untuk mengubah kata-kata yang tidak normal menjadi sesuai dengan standar.
3. *Tokenizing* merupakan tahapan pemecahan kata dari data menjadi bagian-bagian kecil yang disebut token.
4. *Filtering* digunakan untuk menghilangkan kata-kata yang dianggap tidak informatif, proses ini dilakukan sesuai dengan *stopword removal*. *Stopword removal* mengacu pada teks yang tidak memiliki kontribusi signifikan terhadap dokumen.
5. *Stemming* digunakan untuk mengubah kata menjadi kata dasar yang memiliki makna.

2.3. *Topic Modeling* Menggunakan *Latent Dirichlet Allocation*

Topic modeling merupakan salah satu teknik *unsupervised machine learning* untuk menemukan tema tersembunyi dari kumpulan dokumen teks besar guna mengelompokkan tema-tema tersebut menjadi satu

topik [9], [10]. Pemodelan topik menganggap bahwa setiap dokumen merupakan kombinasi dari sekumpulan topik dan kata [19]. *Latent Dirichlet Allocation* (LDA) merupakan salah satu metode dari pemodelan topik yang banyak digunakan. *Latent* mengacu pada segala sesuatu yang tersembunyi dalam data, *Dirichlet* merujuk pada distribusi probabilitas yang menggambarkan distribusi topik dalam dokumen dan kata-kata dalam topik, dan *Allocation* berarti mengalokasikan topik atau kata-kata menjadi lebih spesifik ke dalam dokumen [7]. LDA digunakan untuk menampilkan sebuah topik menggunakan probabilitas dari setiap kata sehingga dapat membantu menggambarkan dokumen-dokumen menjadi lebih terstruktur [11]. *Latent Dirichlet Allocation* dianggap lebih unggul dalam menghasilkan topik bermakna logis [12], kemampuan interpretasi, dan performansi prediktif [13].

Pada tahapan ini, penelitian ini menggunakan dua parameter acuan yaitu *number of topics* dan *words of topic*, untuk menghasilkan model dengan jumlah topik terbaik. Penelitian ini terlebih dahulu menentukan klaster *words of topic* atau jumlah kata per topik, yaitu berjumlah 10 topik, memiliki tujuan agar model yang dibangun menghasilkan model yang jelas dan tidak bersinggungan satu sama lain. Kemudian, penelitian ini menganalisis *number of topics* atau jumlah topik dengan melakukan evaluasi menggunakan nilai *coherence*, dimana kumpulan kata yang dihasilkan oleh topic modeling dinilai berdasarkan tingkat kemudahan pemahaman manusia [20].

Nilai *coherence* dapat mengukur hubungan antar kata dalam pemodelan topik, evaluasi topik terbaik dinilai berdasarkan nilai *coherence* tertinggi yang diperoleh [7]. Di samping itu, penelitian ini juga melakukan evaluasi menggunakan pendekatan *human judgement* untuk menginterpretasikan distribusi kata dari hasil pemodelan topik [17]. Dalam konteks penilaian manusia, penelitian ini menggunakan pendekatan intrusi kata untuk mengevaluasi topik yang dinilai memiliki makna sama dan saling berkaitan satu sama lain sehingga dapat diidentifikasi oleh manusia [21]. Pendekatan ini merujuk pada sejauh mana topik yang dihasilkan oleh model dapat direpresentasikan dengan logis atau alami menurut pandangan manusia. Dalam penelitian ini, penelitian ini memanfaatkan *library* gensim menggunakan software berbasis Python.

2.4. Visualisasi *Topic Modeling*

Hasil analisis *topic modeling* kemudian divisualisasikan menggunakan sistem visualisasi interaktif berbasis web yaitu LDAvis untuk membantu memahami makna setiap topik dan hubungan antara topik yang satu dengan topik lainnya [22]. Penelitian ini menggunakan *library* pyLDAvis pada *software* berbasis Python. Dengan menggunakan PyLDAvis, distribusi kata dan relevansinya dengan masing-masing topik dapat divisualisasikan dengan jelas. Kemudian, hasil visualisasi LDAvis disajikan dalam format .html yang dapat memberikan kemudahan untuk mengeksplorasi hubungan antara topik dan kata-kata yang terbentuk dari topik yang ada.

3. HASIL DAN PEMBAHASAN

3.1. Hasil Pengumpulan Data

Penelitian ini mengumpulkan data menggunakan *search engine hacks* dengan kata kunci Somethinc, misalnya dengan menggabungkan operator site:detik.com intitle:somethinc atau site:detik.com somethinc untuk menampilkan berita terkait merek Somethinc. Rentang waktu yang penelitian ini gunakan mulai dari bulan Januari 2022 sampai dengan Desember 2023. Setelah itu, pengambilan data dilakukan secara manual, data yang penelitian ini ambil adalah isi berita secara menyeluruh., dengan mengecualikan gambar yang terdapat dalam isi berita. Data set disimpan pada file dalam format .xlsx. Total data yang penelitian ini peroleh dari tujuh portal berita adalah 143 data berita mentah. Selengkapnya terdapat dapat dilihat dalam Tabel 1.

Tabel 1. Hasil Pengumpulan Data

Portal Berita	Jumlah
Fimela	35
Kompas	29
Kumparan	24
Marketeers	19
Antara News	15
Detik	12
Tempo	9
Total	143

3.2. Hasil *Preprocessing* Data

Pada tahapan ini, penelitian ini melakukan *preprocessing* data terhadap 143 data berita mentah. Untuk melakukan tahap preprocessing penelitian ini menggunakan bahasa pemrograman pada *software* berbasis Python. Tahapan pertama adalah *case folding*, digunakan untuk mengkonversi huruf kapital menjadi huruf kecil. Setelah itu, dilakukan *cleaning data* untuk menghapus karakter seperti tanda baca, angka, garis miring,

data duplikat dan tidak konsisten. Tahapan kedua yaitu *normalization*, digunakan untuk mengubah kata yang tidak normal menjadi normal. Tahap selanjutnya adalah *tokenizing*, digunakan untuk memisahkan kata pada teks. Tahap keempat yaitu *Filtering*, merupakan tahap eliminasi kata hasil token untuk mewakili dokumen sesuai dengan *stopword removal*. Tahap terakhir yaitu *stemming* digunakan untuk mengubah kata menjadi makna dasar. Setelah data bersih, data kemudian disimpan sebagai ekstensi csv. Adapun contoh dari hasil *preprocessing* data berita dapat dilihat pada Tabel 2.

Tabel 2. Hasil *Preprocessing* Data

Proses	Hasil
Data Collection	Untuk memberikan kesempatan pada semua orang memakai skincare, Somenthinc 'sedekah' produk beauty dalam event Women's Day Out yang digelar 22-23 Oktober 2022 di Sleman City Hall (SCH). Tak hanya skincare, pihaknya juga tawarkan makeup dan body care.
Case Folding	untuk memberikan kesempatan pada semua orang memakai skincare somenthinc sedekah produk beauty dalam event women s day out yang digelar oktober di sleman city hall sch tak hanya skincare pihaknya juga tawarkan makeup dan body care
Normalization	untuk memberikan kesempatan pada semua orang memakai skincare somenthinc sedekah produk beauty dalam event women s day out yang digelar oktober di sleman city hall sch tidak hanya skincare pihaknya juga tawarkan makeup dan body care
Tokenizing	['untuk', 'memberikan', 'kesempatan', 'pada', 'semua', 'orang', 'memakai', 'skincare', 'somenthinc', 'sedekah', 'produk', 'beauty', 'dalam', 'event', 'women', 's', 'day', 'out', 'yang', 'digelar', 'oktober', 'di', 'sleman', 'city', 'hall', 'sch', 'tidak', 'hanya', 'skincare', 'pihaknya', 'juga', 'tawarkan', 'makeup', 'dan', 'body', 'care']
Filtering	['kesempatan', 'orang', 'memakai', 'skincare', 'somenthinc', 'sedekah', 'produk', 'beauty', 'event', 'women', 's', 'day', 'out', 'digelar', 'oktober', 'sleman', 'city', 'hall', 'sch', 'skincare', 'tawarkan', 'makeup', 'body', 'care']
Stemming	['orang', 'pakai', 'skincare', 'somenthinc', 'sedekah', 'produk', 'beauty', 'event', 'women', 's', 'day', 'out', 'gelar', 'oktober', 'sleman', 'city', 'hall', 'sch', 'skincare', 'tawar', 'makeup', 'body', 'care']

3.3. Hasil *Topic Modeling* Menggunakan *Latent Dirichlet Allocation* (LDA)

Pada tahapan *topic modeling* menggunakan *Latent Dirichlet Allocation* (LDA), penelitian ini memanfaatkan *library* gensim yang dirancang untuk mengekstraksi topik semantik secara otomatis dari dokumen-dokumen dengan efisien menggunakan *software* berbasis Python. Penelitian ini terlebih dahulu menentukan klaster *words of topic* yaitu mengacu pada jumlah kata yang akan digunakan untuk menyusun topik. Jumlah kata ditentukan sebanyak 10 kata, dengan tujuan agar topik yang dihasilkan tidak bersinggungan satu sama lain. Selanjutnya penelitian ini melakukan evaluasi pemodelan topik dengan menentukan nilai *coherence*. Parameter yang dijadikan acuan dalam evaluasi ini menggunakan *number of topics* atau jumlah topik yang diperoleh dalam dokumen. Semakin tinggi nilai *coherence*, semakin baik juga pemahaman interpretasi manusia [20] terhadap topik tersebut. Untuk mendapatkan topik dengan model terbaik, penelitian ini menganalisis nilai *coherence* dengan melakukan *running* sebanyak 50 *passes*. Hasil dari nilai *coherence* terdapat pada Tabel 3.

Tabel 3. Hasil Nilai *Coherence*

Topik ke-	Nilai *Coherence*
2	0.34956908503951556
3	0.36215427160146607
4	0.38634393053229040
5	0.40395904371265760
6	0.40426053685302965
7	0.38786463817768285
8	0.33894833506918710
9	0.39859344250239990
10	0.48786738941991264

Berdasarkan informasi yang terdapat pada Tabel 3, menunjukkan bahwa setiap topik memiliki nilai *coherence* yang berbeda. Penelitian ini menggunakan jumlah topik = 6 dan jumlah kata = 10. Model topik 6 memiliki nilai *coherence* sebesar 0.404. Hasil dari pemodelan topik akan direpresentasikan oleh sekumpulan kata yang menggambarkan karakteristik dari masing-masing topik. Kemudian, ditandai dengan label 0, 1, 2, 3, 4, dan 5 berdasarkan nilai probabilitas berupa frekuensi kemunculan kata untuk mewakili topik tertentu. Selanjutnya penelitian ini melakukan evaluasi menggunakan pendekatan penilaian manusia untuk

menginterpretasikan setiap topik berdasarkan kumpulan kata yang sering muncul dalam model yang telah dibuat. Tabel 4 menampilkan hasil evaluasi pemodelan topik dengan mempertimbangkan penilaian manusia.

Tabel 4. Hasil Interpretasi Jumlah Topik 6

Topik	Probabilitas * Kata	Analisis Topik
0	'0.077*"kulit" + 0.032*"wajah" + 0.018*"produk" + 0.017*"somethinc" + 0.017*"skincare" + 0.016*"sunscreen" + 0.015*"kandung" + 0.013*"jerawat" + 0.010*"skin" + 0.010*"bersih"	Produk skincare untuk kulit berjerawat
1	'0.036*"produk" + 0.027*"beauty" + 0.021*"best" + 0.017*"cantik" + 0.016*"brand" + 0.014*"makeup" + 0.012*"tokopedia" + 0.008*"indonesia" + 0.008*"awards" + 0.007*"awat"	Brand kecantikan terbaik di Indonesia yang meraih penghargaan
2	'0.033*"somethinc" + 0.021*"nct" + 0.020*"dream" + 0.019*"hee" + 0.019*"so" + 0.017*"produk" + 0.016*"han" + 0.016*"konsumen" + 0.012*"kolaborasi" + 0.010*"kulit"	Adanya kolaborasi berupa produk diikuti adanya interaksi dengan konsumen
3	'0.043*"somethinc" + 0.041*"serum" + 0.039*"kulit" + 0.022*"warna" + 0.021*"produk" + 0.009*"indonesia" + 0.009*"irene" + 0.008*"lip" + 0.007*"brand" + 0.007*"hadir"	Menghadirkan produk perawatan kulit dan kecantikan
4	'0.046*"produk" + 0.029*"lokal" + 0.028*"brand" + 0.020*"indonesia" + 0.020*"shopee" + 0.014*"somethinc" + 0.009*"umkm" + 0.009*"kampanye" + 0.009*"makeup" + 0.008 *"pasar"	Brand lokal yang melakukan kampanye terkait pemasaran produk
5	'0.043*"somethinc" + 0.025*"produk" + 0.016*"cantik" + 0.016*"brand" + 0.013*"kulit" + 0.010*"lokal" + 0.009*"awat" + 0.008*"indonesia" + 0.008*"skincare" + 0.008*"hadir"	Brand lokal yang hadir dengan produk perawatan kulit atau *skincare*

Melalui proses interpretasi menggunakan pendekatan penilaian manusia seperti terdapat pada Tabel 4 di atas, topik yang paling banyak muncul mengenai merek Somethinc diantaranya meliputi, produk skincare untuk kulit berjerawat, brand kecantikan terbaik di Indonesia yang meraih penghargaan, kolaborasi produk diikuti adanya interaksi dengan konsumen, menghadirkan produk perawatan kulit dan kecantikan, kampanye pemasaran produk, dan brand lokal yang hadir dengan produk perawatan kulit. Temuan tersebut sejalan dengan tujuan utama dari penelitian ini yaitu untuk mengetahui topik yang paling sering dibahas mengenai merek kosmetik Somethinc. Adapun, untuk memastikan validitas hasil penelitian, penelitian ini melakukan evaluasi dengan fokus pada nilai *coherence* tertinggi yaitu jumlah topik 10 dengan nilai *coherence* sebesar 0.489. Hasil pemodelan topik dari jumlah topik 10, dapat dilihat pada Tabel 5.

Tabel 5. Hasil Pemodelan Jumlah Topik 10

Topik	Probabilitas * Kata
0	'0.027*"makeup" + 0.020*"warna" + 0.018*"produk" + 0.015*"somethinc" + 0.014*"douyin" + 0.014*"indonesia" + 0.012*"orang" + 0.011*"glitter" + 0.011*"merch" + 0.010*"ala"'
1	'0.050*"produk" + 0.023*"kulit" + 0.015*"lip" + 0.012*"somethinc" + 0.012*"body" + 0.010*"series" + 0.009*"halal" + 0.009*"brand" + 0.008*"cream" + 0.008*"bibir"'
2	'0.035*"langgan" + 0.028*"produk" + 0.017*"lazada" + 0.014*"butuh" + 0.014*"beauty" + 0.012*"customer" + 0.011*"laz" + 0.011*"pasar" + 0.009*"belanja" + 0.009*"marketing"'
3	'0.078*"kulit" + 0.031*"somethinc" + 0.025*"serum" + 0.023*"wajah" + 0.018*"produk" + 0.018*"kandung" + 0.017*"jerawat" + 0.012*"skincare" + 0.010*"pori" + 0.009*"skin"'
4	'0.055*"warna" + 0.031*"somethinc" + 0.025*"kulit" + 0.023*"makeup" + 0.019*"produk" + 0.019*"cushion" + 0.016*"shade" + 0.016*"undertone" + 0.013*"lip" + 0.012*"tampil"'
5	'0.033*"somethinc" + 0.031*"best" + 0.022*"beauty" + 0.021*"cantik" + 0.019*"tokopedia" + 0.019*"produk" + 0.013*"awat" + 0.011*"awards" + 0.010*"masyarakat" + 0.010*"menang"'
6	'0.037*"sunscreen" + 0.033*"kulit" + 0.026*"wajah" + 0.017*"skincare" + 0.015*"somethinc" + 0.012*"produk" + 0.009*"dream" + 0.009*"libur" + 0.009*"nct" + 0.009*"skin"'
7	'0.046*"somethinc" + 0.040*"serum" + 0.028*"kulit" + 0.026*"produk" + 0.015*"brand" + 0.013*"irene" + 0.012*"indonesia" + 0.011*"lokal" + 0.009*"cantik" + 0.008*"awat"'
8	'0.032*"brand" + 0.027*"produk" + 0.024*"tik" + 0.023*"tok" + 0.016*"indonesia" + 0.014*"lokal" + 0.012*"cantik" + 0.010*"somethinc" + 0.010*"jual" + 0.008*"merek"'
9	'0.040*"produk" + 0.026*"brand" + 0.026*"somethinc" + 0.026*"lokal" + 0.018*"shopee" + 0.015*"indonesia" + 0.012*"hadir" + 0.011*"cantik" + 0.010*"kolaborasi" + 0.010*"konsumen"'

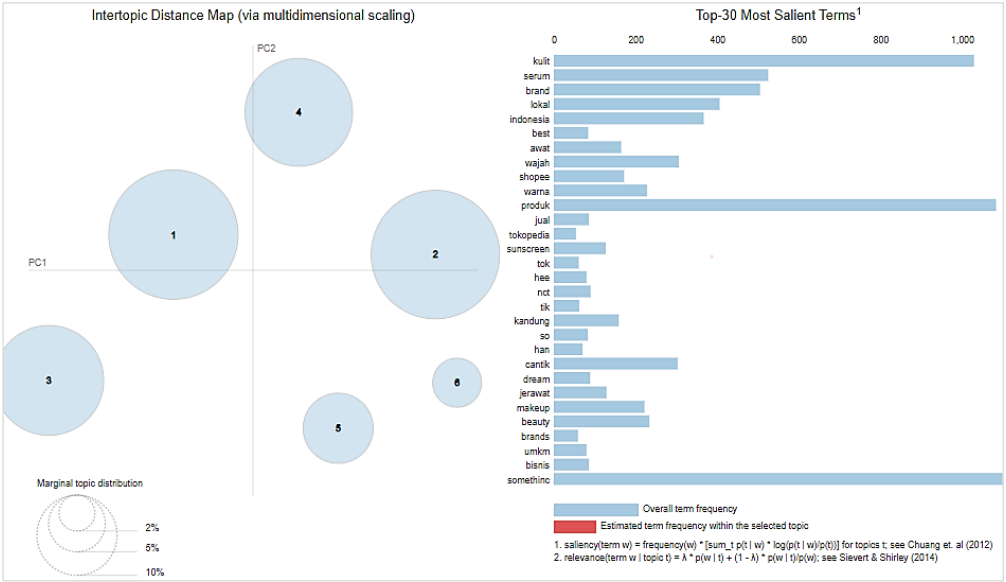
Tabel 5 merupakan persebaran berupa probabilitas kata dari setiap topik untuk jumlah topik 10. Kemudian, penelitian ini melakukan evaluasi pada jumlah topik 10 menggunakan pendekatan penilaian manusia. Dapat diinterpretasikan bahwa, topik 0 membahas tentang tren *makeup* terbaru ala Somethinc, topik 1 membahas brand halal yang memiliki produk dan perawatan kulit dan tubuh, topik 3 membahas produk skincare yaitu serum untuk kulit berjerawat, topik 4 membahas produk makeup Somethinc, topik 5 membahas penghargaan produk kecantikan, topik 6 membahas produk perawatan kulit yaitu *sunscreen*, topik 7 membahas produk perawatan kulit yaitu serum, topik 8 membahas penjualan produk pada *online marketplace*, dan topik 9 membahas tentang kolaborasi produk.

Dari hasil interpretasi yang telah dilakukan oleh penelitian ini, terdapat satu topik yang tidak dapat diinterpretasikan, yaitu terdapat pada topik 2. Topik 2 memiliki kumpulan kata yang terdiri dari langgan, produk, lazada, butuh, *beauty*, *customer*, laz, pasar, belanja, dan *marketing*. Ketika dievaluasi menggunakan penilaian manusia, kumpulan kata tersebut tidak dapat diinterpretasikan karena terdapat beberapa kata yang tidak terkait dengan intrusi kata. Dalam kata-kata *"customer"*, "butuh", "produk", dan *"beauty"* dapat diinterpretasikan bahwa keempat kata tersebut terkait konsumen yang membutuhkan produk kecantikan. Kemudian pada kata seperti "lazada", "laz", dan "belanja" kata-kata tersebut berkaitan dengan online *marketplace*. Disamping itu kata-kata *"marketing"*, "pasar", dan "langgan" berkaitan dengan pemasaran. Hal ini berarti bahwa dalam satu topik terdapat 3 interpretasi yang berbeda. Dalam penilaian manusia, kondisi seperti ini dapat dikatakan buruk sehingga tidak dapat diinterpretasikan [17].

Berdasarkan hasil analisis dan evaluasi, kualitas pembentukan topik terbaik ditentukan berdasarkan hubungan antara kata-kata sesuai dengan topik menggunakan pendekatan penilaian manusia [23]. Ini berarti bahwa hubungan antar kata merupakan hal yang penting. Tidak hanya itu, meskipun suatu topik memiliki nilai *coherence* tertinggi, jika model tersebut memiliki satu atau lebih kata yang tidak dapat diinterpretasikan oleh penilaian manusia, maka model tersebut tetap dianggap buruk [21]. Hal ini menunjukkan bahwa evaluasi menggunakan nilai *coherence* dengan memperhatikan pendekatan penilaian manusia memiliki peran penting dalam menentukan kualitas jumlah topik terbaik. Dalam penelitian ini hasil dengan jumlah topik terbaik terdapat pada jumlah topik 6 dan jumlah kata 10 dengan nilai *coherence* sebesar 0.404. Dapat ditarik kesimpulan bahwa jumlah topik 6 menghasilkan topik-topik yang relevan dalam konteks merek kosmetik Somethinc.

3.4. Hasil Visualisasi Pemodelan Topik

Setelah melakukan pemodelan topik menggunakan *Latent Dirichlet Allocation* (LDA), selanjutnya penelitian ini melakukan visualisasi jumlah topik 6 dengan menggunakan pyLDAvis untuk mengetahui distribusi kata dari setiap topik yang dihasilkan. Gambar 2 di bawah ini menampilkan visualisasi dari pemodelan topik.



Gambar 2. Visualisasi Pemodelan Topik

Berdasarkan Gambar 2 di atas, menunjukan bahwa hasil visualisasi jumlah topik 6 terbagi menjadi dua bagian. Pada bagian kiri terdapat bentuk lingkaran yang digunakan untuk menggambarkan topik dengan menampilkan hubungan antara topik-topik tersebut. Besarnya lingkaran menunjukkan bahwa topik yang

muncul secara berurutan semakin berpengaruh terhadap dokumen. Pada bagian kanan terdapat diagram batang horizontal. Diagram ini digunakan untuk menggambarkan kata-kata yang dominan dibahas dalam topik. Panjang batang mewakili frekuensi topik yang muncul dari satu kata. Bagian kiri dan kanan pada visualisasi tersebut saling terhubung satu sama lain, sederhananya bagian kiri untuk memilih sebuah topik, selanjutnya akan ditunjukkan kata-kata yang paling sering muncul oleh bagian kanan untuk menginterpretasikan topik yang dipilih sebelumnya.

Pada bagian kanan visualisasi yaitu *bar chart*, terdapat 30 kata *(terms)* penting yang menampilkan persentase kemunculan kata untuk masing-masing topik. Kata-kata yang ditampilkan diantaranya yaitu kulit, serum, brand, lokal, Indonesia, best, awat, wajah, shopee, warna, produk, jual, tokopedia, *sunscreen*, tok, hee, nct, tik, kandung, so, han, cantik, dream, jerawat, *makeup*, *beauty*, *brands*, umkm, bisnis, dan somethinc. Dapat disimpulkan bahwa berdasarkan visualisasi menggunakan pyLDAvis, jumlah topik 6 menghasilkan distribusi kata yang relevan terkait merek Somethinc.

4. KESIMPULAN

Berdasarkan hasil penelitian, metode *topic modeling* menggunakan *Latent Dirichlet Allocation* (LDA) dapat dikatakan efektif dalam membentuk topik-topik yang paling sering dibahas pada media berita *online* terkait merek Somethinc. Dalam penelitian ini, jumlah topik terbaik terdapat pada jumlah kata sebanyak 10 topik dan jumlah topik 6 dengan nilai *coherence* sebesar 0.404. Hasil tersebut ditentukan berdasarkan nilai *coherence* tertinggi dengan melakukan running sebanyak 50 *passes* serta memperhatikan pendekatan penilaian manusia yang terbukti berpengaruh dalam menginterpretasikan topik. Interpretasi dari setiap topik mengenai merek Somethinc pada jumlah topik ke-6 diantaranya yaitu, topik 0 membahas produk skincare untuk kulit berjerawat, topik 1 membahas penghargaan brand kecantikan terbaik, topik 2 membahas kolaborasi produk, topik 3 membahas tentang hadirnya produk skincare dan kecantikan, topik 4 membahas kampanye pemasaran produk, dan topik terakhir yaitu 5 membahas brand lokal dengan produk perawatan kecantikan. Di samping itu, visualisasi menggunakan pyLDAvis berhasil memberikan gambaran tentang topik paling sering dibahas dengan menunjukkan persebaran kata yang relevan mengenai merek Somethinc.

Temuan dari penelitian ini dapat dimanfaatkan oleh perusahaan untuk mendapatkan informasi lebih mendalam terkait respon media berita online mengenai merek Somethinc, sehingga perusahaan dapat meningkatkan kualitasnya di masa mendatang. Selain itu, masyarakat juga dapat memperoleh informasi yang relevan mengenai merek Somethinc. Adapun keterbatasan dalam penelitian ini yaitu pengumpulan data berita dilakukan secara manual, sehingga data yang diperoleh menjadi terbatas. Saran untuk penelitian selanjutnya peneliti dapat memanfaatkan teknik *scraping* menggunakan perangkat lunak *(software)* ataupun bahasa pemrograman untuk mendapatkan volume data yang lebih besar. Evaluasi pemodelan topik menggunakan pendekatan penilaian manusia perlu dikaji lebih mendalam khusunya oleh seorang ahli untuk mendapatkan hasil *topic modeling* yang lebih akurat.

REFERENSI

[1] Kementerian Perindustrian Indonesia, "Rencana Induk Pembangunan Industri Nasional 2015 - 2035," 2015. [Online]. Available: https://kemenperin.go.id/ripin.pdf

[2] Kementerian Perindustrian Republik Indonesia, "Perkembangan Industri Kosmetik Nasional." http://ikft.kemenperin.go.id/perkembangan-industri-kosmetik-nasional/ (accessed Nov. 15, 2023).

[3] N. Amberg and C. Fogarassy, "Green consumer behavior in the cosmetics market," *Resources*, vol. 8, no. 3, pp. 1–19, 2019, doi: 10.3390/resources8030137.

[4] M. Ferdinand and W. S. Ciptono, "Indonesia's Cosmetics Industry Attractiveness, Competitiveness and Critical Success Factor Analysis," *J. Manaj. Teor. dan Terap. | J. Theory Appl. Manag.*, vol. 15, no. 2, pp. 209–223, 2022, doi: 10.20473/jmtt.v15i2.37451.

[5] M. Arifiansyah Ayub, "Analisis Topik Ekonomi Dengan Algoritma K-Means Pada Media Online Era Pandemi Covid-19 Di Sulawesi Tenggara," *JIKO (Jurnal Inform. dan Komputer)*, vol. 4, no. 2, pp. 133–138, 2021, doi: 10.33387/jiko.v4i2.3235.

[6] S. Kurniawan, W. Gata, D. A. Puspitawati, N. -, M. Tabrani, and K. Novel, "Perbandingan Metode Klasifikasi Analisis Sentimen Tokoh Politik Pada Komentar Media Berita Online," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 2, pp. 176–183, 2019, doi: 10.29207/resti.v3i2.935.

[7] M. D. R Wahyudi, A. Fatwanto, U. Kiftiyani, and M. Galih Wonoseto, "Topic Modeling of Online Media News Titles during COVID-19 Emergency Response in Indonesia Using the Latent Dirichlet Allocation (LDA) Algorithm," *Telematika*, vol. 14, no. 2, pp. 101–111, 2021, doi: 10.35671/telematika.v14i2.1225.

[8] C. Naury, D. H. Fudholi, and A. F. Hidayatullah, "Topic Modelling pada Sentimen Terhadap Headline Berita Online Berbahasa Indonesia Menggunakan LDA dan LSTM," *J. Media Inform. Budidarma*, vol. 5, no. 1, pp. 24–33, 2021, doi: 10.30865/mib.v5i1.2556.

[9] P. Kherwa and P. Bansal, "Topic Modeling: A Comprehensive Review," *EAI Endorsed Trans. Scalable Inf. Syst.*, vol. 7, no. 24, pp. 1–16, 2020, doi: 10.4108/eai.13-7-2018.159623.

[10] A. H. Marani and E. P. S. Baumer, "A Review of Stability in Topic Modeling: Metrics for Assessing and Techniques for Improving Stability," *ACM Comput. Surv.*, 2023, doi: 10.1145/3623269.
[11] A. Farkhod, A. Abdusalomov, F. Makhmudov, and Y. I. Cho, "LDA-Based Topic Modeling Sentiment Analysis Using Topic/Document/Sentence (TDS) Model," *Appl. Sci.*, vol. 11, no. 23, pp. 1–15, 2021, doi: 10.3390/app112311091.
[12] R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis," *Front. Artif. Intell.*, vol. 3, no. July, pp. 1–14, 2020, doi: 10.3389/frai.2020.00042.
[13] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic Modeling in Embedding Spaces," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 439–453, 2020, doi: 10.1162/tacl_a_00325.
[14] W. A. Nawang Sari and H. Dwi Purnomo, "Topic Modeling Using the Latent Dirichlet Allocation Method on Wikipedia Pandemic Covid-19 Data in Indonesia," *J. Tek. Inform.*, vol. 3, no. 5, pp. 1223–1230, 2022, doi: 10.20884/1.jutif.2022.3.5.321.
[15] P. Patmawati and M. Yusuf, "Analisis Topik Modelling Terhadap Penggunaan Sosial Media Twitter oleh Pejabat Negara," *Build. Informatics, Technol. Sci.*, vol. 3, no. 3, pp. 122–129, 2021, doi: 10.47065/bits.v3i3.1012.
[16] K. R. A. P. Santoso, A. Husna, N. W. Putri, and N. A. Rakhmawati, "Analisis Topik Tagar Covidindonesia pada Instagram Menggunakan Latent Dirichlet Allocation," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 7, no. 1, pp. 1–9, 2022, doi: 10.14421/jiska.2022.7.1.1-9.
[17] Wiranto and M. R. Uswatunnisa, "Topic Modeling for Support Ticket using Latent Dirichlet Allocation," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 6, pp. 998–1005, 2022, doi: 10.29207/resti.v6i6.4542.
[18] E. A. Felix and S. P. Lee, "Systematic literature review of preprocessing techniques for imbalanced data," *IET Softw.*, vol. 13, no. 6, pp. 479–496, 2019, doi: 10.1049/iet-sen.2018.5193.
[19] A. M. Grisales, S. Robledo, and M. Zuluaga, "Topic Modeling: Perspectives From a Literature Review," *IEEE Access*, vol. 11, no. January, pp. 4066–4078, 2023, doi: 10.1109/ACCESS.2022.3232939.
[20] M. R. A. Fahlevvi, "Topic Modeling on Online News Portal Using Latent Dirichlet Allocation (LDA)," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 16, no. 4, pp. 335–344, 2022, doi: 10.22146/ijccs.74383.
[21] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models," in *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*, 2009, pp. 288–296.
[22] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014, pp. 63–70. doi: 10.3115/v1/w14-3110.
[23] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proceeding of the National Academy of Sciences*, 2004, pp. 5228–5235. doi: 10.1073/pnas.0307752101.