TOPIC MODELING KOMENTAR PRODUK *SKINCARE* DAN PEMBUATAN ONTOLOGINYA UNTUK REPRESENTASI PENGETAHUAN ULASAN PRODUK PEMENANG *BEST OF BEAUTY AWARDS* PADA FEMALE DAILY



Disusun Oleh:

Nama : Nabila
NIM : 21523216

PROGRAM STUDI INFORMATIKA – PROGRAM SARJANA
FAKULTAS TEKNOLOGI INDUSTRI
UNIVERSITAS ISLAM INDONESIA
2025

HALAMAN PENGESAHAN DOSEN PEMBIMBING

TOPIC MODELING KOMENTAR PRODUK *SKINCARE* DAN PEMBUATAN ONTOLOGINYA UNTUK REPRESENTASI PENGETAHUAN ULASAN PRODUK PEMENANG *BEST OF BEAUTY AWARDS* PADA FEMALE DAILY

TUGAS AKHIR

Disusun Oleh:

Nama : Nabila

NIM : 21523216

Yogyakarta, 30 Juli 2025

Pembimbing,

(Chanifah Indah Ratnasari, S.Kom, M.Kom.)

HALAMAN PENGESAHAN DOSEN PENGUJI

TOPIC MODELING KOMENTAR PRODUK *SKINCARE* DAN PEMBUATAN ONTOLOGINYA UNTUK REPRESENTASI PENGETAHUAN ULASAN PRODUK PEMENANG *BEST OF BEAUTY AWARDS* PADA FEMALE DAILY

TUGAS AKHIR

Telah dipertahankan di depan sidang penguji sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer dari Program Studi Informatika – Program Sarjana di Fakultas Teknologi Industri Universitas Islam Indonesia

Yogyakarta, 23 Juli 2025

Tim Penguji

Chanifah Indah Ratnasari, S.Kom, M.Kom. _____

Anggota 1

Dr. Yudi Prayudi, S.Si., M.Kom. _____

Anggota 2

Elyza Gustri Wahyuni, S.T., M.Cs. _____

Mengetahui,

Ketua Program Studi Informatika – Program Sarjana

Fakultas Teknologi Industri

Universitas Islam Indonesia

(Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D.)

HALAMAN PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan di bawah ini:

Nama : Nabila

NIM : 21523216

Tugas akhir dengan judul:

TOPIC MODELING KOMENTAR PRODUK *SKINCARE* DAN PEMBUATAN ONTOLOGINYA UNTUK REPRESENTASI PENGETAHUAN ULASAN PRODUK PEMENANG *BEST OF BEAUTY AWARDS* PADA FEMALE DAILY

Menyatakan bahwa seluruh komponen dan isi dalam tugas akhir ini adalah hasil karya saya sendiri. Apabila di kemudian hari terbukti ada beberapa bagian dari karya ini adalah bukan hasil karya sendiri, tugas akhir yang diajukan sebagai hasil karya sendiri ini siap ditarik kembali dan siap menanggung risiko dan konsekuensi apapun.

Demikian surat pernyataan ini dibuat, semoga dapat dipergunakan sebagaimana mestinya.

Yogyakarta, 30 Juli 2025



(Nabila)

HALAMAN PERSEMBAHAN

Tugas akhir ini dengan sepenuh hati saya persembahkan untuk keluarga tercinta, papa, almarhumah mama, tante, dan ketiga kakak saya yang tak pernah lelah memberikan cinta, doa, dukungan, dan semangat di setiap langkah perjuangan saya hingga titik ini. Tanpa mereka, pencapaian ini tidak akan pernah menjadi nyata. Ucapan terima kasih yang tulus juga saya haturkan kepada sahabat-sahabat terdekat saya, Tirtir, yang selalu hadir dengan tawa, perhatian, dan dorongan penuh makna selama proses panjang meraih gelar sarjana ini. Tak lupa, untuk Andika Raka, yang tak hanya menjadi teman seperjuangan, namun juga saksi bisu setiap tetes keringat dan perjuangan di balik layar. Dan kepada seluruh teman-teman seperjuangan Informatika angkatan 21, yang telah mengukir kenangan tak terlupakan dan menjadi bagian tak terpisahkan dari setiap langkah perjalanan ini. Kalian semua adalah bagian dari keberhasilan ini.

HALAMAN MOTO

"Maka, sesungguhnya beserta kesulitan ada kemudahan"
(Q.S. Al-Insyirah ayat 5)

"*Get a diploma*."
(S.coups of Seventeen)

"Skripsi yang baik adalah skripsi yang selesai."
(Unknown)

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh,

Alhamdulillah, segala puji dan syukur yang tak terhingga penulis panjatkan ke hadirat Allah Swt., atas limpahan rahmat, kasih sayang, serta kekuatan yang senantiasa mengiringi setiap langkah hingga akhirnya karya ilmiah ini, tugas akhir berjudul "*Topic Modeling* Komentar Produk *Skincare* dan Pembuatan Ontologinya untuk Representasi Pengetahuan Ulasan Produk Pemenang *Best Of Beauty Awards* pada Female Daily", dapat diselesaikan dengan baik. Skripsi ini bukan sekadar hasil dari proses akademik, tetapi juga perjalanan panjang yang penuh perjuangan, doa, keraguan, semangat, air mata, dan cinta. Setiap halaman yang tertulis adalah bukti bahwa penulis tidak pernah sendiri. Di balik lembar-lembar ini, ada tangan-tangan yang menopang dan hati-hati yang mendoakan.

Dengan tulus dan penuh rasa hormat, penulis mengucapkan terima kasih yang sedalam-dalamnya kepada:

1. Bapak Prof. Fathul Wahid, S.T.,M.Sc.,Ph.D selaku Rektor Universitas Islam Indonesia.
2. Bapak Hari Purnomo, Prof., Dr., Ir., M.T., IPU, ASEAN.Eng selaku Dekan Fakultas Teknologi Industri.
3. Bapak Dr. Raden Teduh Dirgahayu, S.T., M.Sc. selaku Ketua Jurusan Informatika.
4. Bapak Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D. selaku Kepala Program Studi Informatika.
5. Ibu Chanifah Indah Ratnasari, S.Kom., M.Kom. selaku dosen pembimbing yang dengan sabar, bijak, dan penuh perhatian membimbing penulis selama penyusunan skripsi ini.
6. Papa dan tante selaku orang tua yang selalu mendukung secara materi maupun nonmateri, doa yang selalu mengiringi, dan semangat sehingga tugas akhir ini dapat diselesaikan dengan baik.
7. Almarhumah mama tercinta, terima kasih atas cinta tanpa syarat dan doa yang tak pernah putus. Meskipun ragamu tak lagi mendampingi, setiap langkah dalam perjuangan ini adalah bayang doamu yang masih setia menyertai. Gelar ini adalah bagian dari janji yang ingin kutuntaskan untukmu. Semoga setiap tetes usaha ini sampai sebagai hadiah untukmu di keabadian. Al-Fatihah.
8. Fatimah, Aisyah, dan Atikah selaku kakak yang selalu memberi dukungan secara materi maupun nonmateri, selalu hadir dengan dukungan dan semangat sehingga

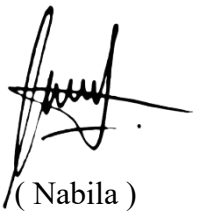
tugas akhir ini dapat diselesaikan dengan baik. yang selalu hadir dengan dukungan dan semangat.

9. Sigita, Wawa, Erika, Caca, Hawa, dan Nana dari Tirtir selaku sahabat saya selalu menemani, menghibur dan memberi dukungan sehingga tugas akhir ini dapat diselesaikan dengan baik. Kalian bagian dari perjalanan ini.
10. Andika Raka dan teman-teman seperjuangan Informatika angkatan 21 yang telah menjadi rumah kedua selama masa studi. Terima kasih untuk tawa, perjuangan, dan kenangan yang tak tergantikan.
11. Dan terakhir, untuk diri saya sendiri, Nabila. Terima kasih telah bertahan. Terima kasih sudah tidak menyerah meski dunia terasa berat. Terima kasih karena terus percaya, melangkah, dan tumbuh. Kamu layak sampai di sini.

Penulis menyadari bahwa karya ini masih jauh dari sempurna. Namun besar harapan penulis, semoga skripsi ini dapat memberi manfaat, baik bagi diri penulis maupun bagi siapa saja yang membaca dan membutuhkan. Akhir kata, semoga Allah Swt. senantiasa melimpahkan rahmat dan hidayah-Nya kepada kita semua, serta memberkahi setiap ilmu yang kita perjuangkan. Aamiin ya Rabbal ‘Alamin.

Wassalamualaikum *Warahmatullahi Wabarakatuh.*

Yogyakarta, 30 Juli 2025



(Nabila)

SARI

Meningkatnya popularitas produk perawatan kulit di Indonesia telah mendorong banyak konsumen untuk mencari dan berbagi informasi melalui platform daring. Salah satu platform yang paling berpengaruh adalah Female Daily, yang menyediakan ruang bagi pengguna untuk mengulas dan mendiskusikan berbagai produk *skincare*. Penelitian ini bertujuan untuk mengeksplorasi topik-topik dominan dalam komentar pengguna terkait produk *skincare* di Female Daily. Penelitian ini menggunakan pendekatan deskriptif kualitatif dengan metode *topic modeling* berbasis *Latent Dirichlet Allocation* (LDA). Data diperoleh dari komentar pengguna terhadap sejumlah produk *skincare* populer dan diproses melalui tahapan *preprocessing* seperti penghapusan tanda baca, *case folding*, tokenisasi, normalisasi, penghapusan *stopword*, dan *stemming*. Jumlah topik optimal ditentukan berdasarkan skor koherensi. Hasil penelitian menunjukkan bahwa pengguna sering mendiskusikan pengalaman pribadi, menyoroti manfaat dan kekurangan produk, serta merujuk pada permasalahan kulit yang spesifik. Temuan ini menunjukkan bahwa LDA efektif untuk mengekstraksi struktur tema dari korpus ulasan pengguna terkait produk *skincare* pada Female Daily. Hasil ini kemudian dipetakan menjadi ontologi menggunakan Protégé (OWL), meliputi perancangan kelas inti, atribut, dan relasi konseptual yang merepresentasikan entitas dan keterhubungannya di domain *skincare* pada Female Daily. Validasi ontologi dilakukan melalui *query* SPARQL untuk menguji keterhubungan antar-entitas serta konsistensi pengetahuan yang diturunkan dari model topik dengan serangkaian pertanyaan relevan diajukan sebagai skenario pengujian. Ontologi yang dibangun memperkaya hasil pemodelan dengan kerangka semantik yang eksplisit, sehingga memudahkan penelusuran berbasis makna dan pengolahan pengetahuan lanjutan. Kombinasi *topic modeling* dan ontologi ini mampu menyajikan pengetahuan secara terstruktur dan semantik, sehingga mempermudah pengguna dalam menemukan produk skincare yang sesuai dengan kebutuhannya pada domain skincare di platform Female Daily.

Kata kunci: Female Daily, Komentar, LDA, Ontologi, *Skincare*, *Topic Modeling*.

GLOSARIUM

Cohorence Score Digunakan sebagai metode evaluasi topik melalui pengukuran kedekatan semantik antara kata-kata dalam satu topik.

Gensim Sebuah *library* pada bahasa pemrograman Python yang dirancang untuk menangani berbagai tugas terkait pemrosesan bahasa alami (NLP), seperti tokenisasi, analisis sintaksis, dan pemodelan teks.

Ontologi Landasan dalam web semantik untuk menghubungkan pemahaman konsep manusia dengan representasi data yang dapat diproses komputer.

Preprocessing Proses untuk membersihkan data mentah yang tidak sempurna, tidak konsisten, atau bermasalah sebelum informasi yang relevan diekstrak dan dikonversi ke dalam format terstruktur untuk analisis lebih lanjut.

Scraping Data Teknik pengambilan data tertentu secara semi-terstruktur pada sebuah halaman web.

Topic Modeling Pendekatan untuk menemukan topik tersembunyi dalam kumpulan dokumen.

DAFTAR ISI

HALAMAN JUDUL i
HALAMAN PENGESAHAN DOSEN PEMBIMBING ii
HALAMAN PENGESAHAN DOSEN PENGUJI iii
HALAMAN PERNYATAAN KEASLIAN TUGAS AKHIR iv
HALAMAN PERSEMBAHAN v
HALAMAN MOTO vi
KATA PENGANTAR vii
SARI ix
GLOSARIUM x
DAFTAR ISI xi
DAFTAR TABEL xiii
DAFTAR GAMBAR xv
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 3
1.3 Batasan Masalah 3
1.4 Tujuan Penelitian 4
1.5 Manfaat Penelitian 4
1.6 Sistematika Penulisan 4
BAB II LANDASAN TEORI 6
2.1 Female Daily 6
2.2 *Electronic Word of Mouth* (e-WOM) 8
2.3 *Text Mining* 8
2.4 *Topic Modeling* 8
2.5 Preprocessing 11
2.6 *Term Frequency-Inverse Document Frequency* (TF-IDF) 12
2.7 *Latent Dirichlet Allocation* (LDA) 13
2.8 Web Semantik 16
2.9 Ontologi 17
2.10 *Ontology Web Language* (OWL) 17
2.11 SPARQL 18
2.12 Protégé 18
2.13 Python 18
2.14 Penelitian Terdahulu 19
BAB III METODE PENELITIAN 27
3.1 Tahapan Penelitian 27
3.2 Pengumpulan Data 28
3.3 *Preprocessing* Data 29
3.4 *Term Frequency-Inverse Document Frequency* (TF-IDF) 30
3.5 *Topic Modeling* Menggunakan *Latent Dirichlet Allocation* (LDA) 31
3.6 Analisis Hasil 31

3.7 Pembuatan Ontologi....31
3.8 Evaluasi Ontologi....31
BAB IV HASIL DAN PEMBAHASAN....32
4.1 Pengumpulan Data....32
4.2 Implementasi *Preprocessing*....35
4.3 Identifikasi Frasa....47
4.4 Pembobotan dengan TF-IDF....48
4.5 *Topic Modeling* dengan LDA....49
4.6 Hasil dan Analisis....64
4.7 Pembuatan Ontologi....73
BAB V KESIMPULAN DAN SARAN....90
5.1 Kesimpulan....90
5.2 Saran....91
DAFTAR PUSTAKA....92
LAMPIRAN....98

DAFTAR TABEL

Tabel 2.1 Bingkai Analisis Berdasarkan Metode *Topic Modeling* (Bagian 1) 19
Tabel 2.1 Bingkai Analisis Berdasarkan Metode *Topic Modeling* (Bagian 2) 20
Tabel 2.1 Bingkai Analisis Berdasarkan Metode *Topic Modeling* (Bagian 3) 21
Tabel 2.2 Kelebihan dan Kekurangan Metode LDA, LSA, dan PLSA (Bagian 1) 22
Tabel 2.2 Kelebihan dan Kekurangan Metode LDA, LSA, dan PLSA (Bagian 2) 23
Tabel 2.3 Bingkai Analisis Berdasarkan Pembuatan Ontologi (Bagian 1) 24
Tabel 2.3 Bingkai Analisis Berdasarkan Pembuatan Ontologi (Bagian 2) 25
Tabel 4.1 Hasil *Scraping* Data 35
Tabel 4.2 Hasil Data Setelah Dilakukan *Remove Punctuation* (Bagian 1) 37
Tabel 4.2 Hasil Data Setelah Dilakukan *Remove Punctuation* (Bagian 2) 38
Tabel 4.3 Hasil Data Setelah Dilakukan *Case Folding* 38
Tabel 4.4 Hasil Data Setelah Dilakukan *Tokenizing* 39
Tabel 4.5 Contoh Singkatan dan Perbaikannya 41
Tabel 4.6 Contoh Kata *Slang* dan Perbaikannya 41
Tabel 4.7 Contoh Kata Yang Tidak Sesuai Standar (*Typo*) dan Perbaikannya 41
Tabel 4.8 Contoh Kata Dengan Karakter Berulang dan Perbaikannya 42
Tabel 4.9 Hasil Data Setelah Dilakukan *Normalization* 42
Tabel 4.10 Hasil Data Setelah Dilakukan *Stopwords* Awal 43
Tabel 4.11 Hasil Data Setelah Dilakukan *Stopwords* Final 44
Tabel 4.12 Hasil Data Setelah Dilakukan *Stopwords* 46
Tabel 4.13 Hasil Data Setelah Dilakukan Identifikasi Frasa 48
Tabel 4.14 *Coherence Value* (Bagian 1) 51
Tabel 4.14 *Coherence Value* (Bagian 2) 52
Tabel 4.15 Contoh Komentar yang Relevan dengan Topik 1 55
Tabel 4.16 Contoh Komentar yang Relevan dengan Topik 2 56
Tabel 4.17 Contoh Komentar yang Relevan dengan Topik 3 57
Tabel 4.18 Contoh Komentar yang Relevan dengan Topik 4 58
Tabel 4.19 Contoh Komentar yang Relevan dengan Topik 5 59
Tabel 4.20 Contoh Komentar yang Relevan dengan Topik 6 60
Tabel 4.21 Contoh Komentar yang Relevan dengan Topik 7 61
Tabel 4.22 Analisis Hasil *Topic Modeling* 64

Tabel 4.23 Hasil Validasi Evaluator (Bagian 1) 68
Tabel 4.23 Hasil Validasi Evaluator (Bagian 2) 69
Tabel 4.23 Hasil Validasi Evaluator (Bagian 3) 70
Tabel 4.23 Hasil Validasi Evaluator (Bagian 4) 71
Tabel 4.23 Hasil Validasi Evaluator (Bagian 5) 72
Tabel 4.23 Ontologi Spesifikasi Dokumen 76
Tabel 4.24 *Object Property* dan Relasi Antar Konsep (Bagian 1) 78
Tabel 4.24 *Object Property* dan Relasi Antar Konsep (Bagian 2) 79
Tabel 4.25 Definisi *Property* Data 80
Tabel 4.26 Pertanyaan Pengujian Ontologi *Skincare* 85

DAFTAR GAMBAR

Gambar 2.1.1 Logo Female Daily 6
Gambar 2.2 Ilustrasi LDA 14
Gambar 2.3 Visualisasi Grafik Model LDA 15
Gambar 3.1 Tahapan Penelitian 27
Gambar 3.2 Alur Tahapan *Preprocessing* 29
Gambar 4.1 *Import library* Python 32
Gambar 4.2 Inisialisasi Variabel dan *Loop* untuk Mengambil Data 32
Gambar 4.3 Penyimpanan Data ke dalam Format .xlsx 34
Gambar 4.4 Kode Program *Remove Punctuation* 37
Gambar 4.5 Kode Program *Tokenizing* 39
Gambar 4.6 Kode Program *Normalization* 40
Gambar 4.7 Kode Program *Stopwords* 45
Gambar 4.8 Kode Program *Stemming* 46
Gambar 4.9 Kode Program Algoritma Bigram dan Trigram 47
Gambar 4.10 Kode Program Pembobotan TF-IDF 49
Gambar 4.11 Kode Program Menghitung *Coherence Score* 50
Gambar 4.12 Grafik Limit 10 50
Gambar 4.13 Grafik Limit 15 50
Gambar 4.15 Kode Program Visualisasi Model LDA 52
Gambar 4.16 Visualisasi LDA dengan Limit 10 52
Gambar 4.17 Visualisasi LDA dengan Limit 15 53
Gambar 4.18 Visualisasi LDA dengan Limit 20 53
Gambar 4.19 Visualisasi LDA dengan Limit 10 54
Gambar 4.20 Visualisasi *topic*-1 55
Gambar 4.21 Visualisasi *topic*-2 56
Gambar 4.22 Visualisasi *topic*-3 57
Gambar 4.23 Visualisasi *topic*-4 58
Gambar 4.24 Visualisasi *topic*-5 59
Gambar 4.25 Visualisasi *topic*-6 60
Gambar 4.26 Visualisasi *topic*-7 61
Gambar 4.27 *Worldcloud 1* 62

Gambar 4.28 *Worldcloud 2*.....62
Gambar 4.29 *Worldcloud 3*.....62
Gambar 4.30 *Worldcloud 4*.....62
Gambar 4.31 *Worldcloud 5*.....62
Gambar 4.32 *Worldcloud 6*.....62
Gambar 4.34 Implementasi *Worldcloud*.....63
Gambar 4.35 Alur Pembuatan Ontologi 68
Gambar 4.36 Alur Pembuatan Ontologi 74
Gambar 4.37 Diagram Konseptual Relasi Antar *Class*.....77
Gambar 4.38 Struktur *Class* Ontologi 82
Gambar 4.39 *Object Property* Ontologi.....82
Gambar 4.40 *Instances* Ontologi 83
Gambar 4.41 Visualisasi dengan Ontograf.....84
Gambar 4.42 Hasil Query Pertanyaan 1 85
Gambar 4.43 Hasil Query Pertanyaan 2 86
Gambar 4.44 Hasil Query Pertanyaan 3 87
Gambar 4.45 Hasil Query Pertanyaan 4 88

BAB I
PENDAHULUAN

1.1 Latar Belakang

Kosmetik didefinisikan sebagai bahan atau sediaan yang ditujukan untuk perawatan eksternal tubuh manusia yaitu kulit, rambut, kuku, bibir, area genital eksternal, gigi dan membran mukosa mulut. Produk ini dirancang dengan manfaat untuk membersihkan, merawat, mengharumkan, mengubah penampilan, dan/atau memperbaiki bau badan atau melindungi atau memelihara tubuh pada kondisi baik (Peraturan Menteri Kesehatan, 2010). Industri kosmetik diidentifikasi oleh Kementerian Perindustrian Indonesia sebagai salah satu penggerak utama perekonomian nasional. Berdasarkan keterangan dari Waluyo (2023), industri kosmetik di Indonesia mengalami pertumbuhan yang sangat pesat mencapai 21,9%, yakni terdapat 913 perusahaan pada tahun 2022 dan pada pertengahan tahun 2023 terdapat 1.010 perusahaan. Industri kosmetik di Indonesia terus berinovasi dengan menghadirkan beragam produk kecantikan, salah satunya yaitu kosmetik perawatan kulit atau *skincare.* Dengan nilai pasar mencapai USD 2,05 miliar pada tahun 2022, produk *skincare* menjadi salah satu segmen yang tumbuh paling pesat dalam industri perawatan diri di Indonesia. Meskipun produk perawatan pribadi (*personal care*) secara keseluruhan mencapai nilai yang lebih tinggi yaitu USD 3,18 miliar, namun pertumbuhan *skincare* menunjukkan potensi yang sangat besar (Kementerian Koordinator Bidang Perekonomian Republik Indonesia, 2024). Persaingan antar perusahaan juga semakin ketat seiring kemunculan merek-merek baru yang terus berinovasi untuk menarik konsumen (Khotimmah dkk., 2024).

Di era digital, keputusan pembelian konsumen, terutama untuk produk kosmetik, sangat dipengaruhi oleh ulasan *online*. Baik ulasan positif maupun negatif memiliki peran penting dalam membentuk persepsi konsumen terhadap suatu produk (Elwalda dkk., 2016). Temuan Thoumrungroje (2014) juga mendukung hal ini, menunjukkan bahwa konsumen cenderung mengandalkan pengalaman pribadi yang dibagikan di media sosial. Kemudahan akses internet melalui perangkat *mobile* membuat konsumen semakin aktif mencari informasi dan rekomendasi dari pengguna lain (Amin dkk., 2014). Umumnya penilaian konsumen *online* pada suatu produk berupa komentar dan penilaian bintang. Semakin tinggi jumlah bintang yang diberikan, maka semakin positif persepsi konsumen terhadap produk yang dibeli. Suatu produk, khususnya *skincare* yang akan dibeli perlu untuk diketahui bagaimana kualitasnya dengan melihat dari pengalaman atau ulasan para konsumen yang sudah mencoba produk

tersebut sebelumnya sebagai referensi. Hal ini penting untuk dilakukan karena variasi tipe kulit manusia yang luas serta perbedaan individu dalam mengevaluasi kualitas dan kecocokan suatu produk. Ulasan ini dapat ditemukan di kolom komentar produk atau pada forum kecantikan, salah satunya pada platform Female Daily.

Selain Female Daily, terdapat platform kecantikan lain di Indonesia yang serupa, seperti Sociolla dan Beautyhaul yang juga menyediakan fitur ulasan produk. Namun, kelengkapan informasi yang ditampilkan berbeda. Female Daily menyajikan data yang lebih komprehensif, mencakup nama *reviewer*, *rating*, umur, tipe kulit, warna kulit, jangka pemakaian, tempat pembelian produk yang diulas, dan komentar. Sociolla hanya menampilkan nama *reviewer*, *rating*, dan ukuran produk yang diulas, sedangkan Beautyhaul memuat nama *reviewer*, umur, *rating*, tipe kulit, dan permasalahan kulit. Karena penelitian ini berfokus pada analisis komentar pengguna untuk penerapan *topic modeling*, kelengkapan atribut ulasan menjadi penting untuk memberikan konteks yang kaya. Oleh karena itu, Female Daily dipilih sebagai sumber data utama karena menyediakan informasi paling detail dan relevan untuk mendukung analisis.

Female Daily merupakan platform sekaligus forum *online* untuk berbagi informasi, pengalaman, dan rekomendasi mengenai produk-produk kecantikan serta perawatan kulit (Wahyu & Putri, 2023). Female Daily memiliki laman *website* dengan domain femaledaily.com dan versi aplikasi *mobile* bernama 'Female Daily - Beauty Review' dengan jumlah 1 juta unduhan dan jumlah member pengguna yang terdaftar lebih dari 990 ribu. Female Daily mencatat lebih dari 2,5 juta ulasan produk oleh pengguna, dengan lebih dari 100.000 produk telah diulas. Di antara berbagai kategori, *skincare* menjadi kategori paling dominan dengan total ulasan melebihi 1,35 juta (Female Daily, t.t.) Selain itu, platform ini juga menawarkan fitur-fitur menarik, seperti ulasan produk (FD *Review),* konten editorial berisi artikel-artikel (FD Editorial), *E-Commerce* (FD Studio) dan forum diskusi (FD *Talk*). Pada fitur FD *Review*, pengguna dapat mencari produk yang ingin dibeli dan dapat membaca informasi tentang ulasan, tipe kulit, jangka pemakaian, dan penilaian bintang (*rate*) dari konsumen lain mengenai sebuah produk yang telah mereka gunakan.

Ulasan-ulasan yang telah ditulis oleh konsumen tidak hanya memberikan informasi mengenai produk, tetapi juga mencerminkan pengalaman pengguna secara langsung. Namun, banyaknya jumlah dan beragam ulasan membuat pengguna kesulitan untuk menemukan informasi sebuah produk *skincare* yang relevan untuk permasalahan kulit pengguna. Kompleksitas informasi ini membuat konsumen membutuhkan panduan untuk memilih produk yang sesuai dengan kebutuhan mereka.

Membaca seluruh ulasan yang tersedia akan memakan banyak waktu dan tenaga, sehingga dibutuhkan pendekatan analisis teks yang mampu mengelompokkan komentar berdasarkan kesamaan topik secara otomatis. Salah satu pendekatan yang banyak digunakan dalam analisis teks tidak terstruktur adalah *topic modeling*, yaitu metode yang dapat mengekstraksi tema-tema utama dari sekumpulan dokumen. Dalam penelitian ini, *topic modeling* diterapkan untuk mengidentifikasi topik-topik dominan pada komentar produk skincare yang memperoleh penghargaan ***Best of Beauty Awards*** di platform Female Daily dengan menggunakan metode *Latent Dirichlet Allocation* (LDA). Metode LDA dipilih karena bersifat generatif, probabilistik, mampu menangani korpus besar tanpa pelabelan manual, dan terbukti menghasilkan topik yang koheren. Selanjutnya, hasil analisis topik tersebut diformalisasikan ke dalam sebuah ontologi domain *skincare* yang merepresentasikan konsep, relasi, dan kata kunci yang sering digunakan oleh pengguna Female Daily. Ontologi yang dihasilkan bertujuan untuk menyajikan pengetahuan secara terstruktur dan semantik, sehingga mempermudah pengguna dalam menemukan produk *skincare* yang sesuai dengan kebutuhannya pada domain *skincare* di platform Female Daily.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, maka penelitian ini akan memfokuskan pada masalah berikut:

a. Topik utama apa saja yang muncul dalam komentar produk s*kincare* di Female Daily menggunakan teknik *topic modeling* dengan mengimplementasikan metode *Latent Dirichlet Allocation* (LDA)*?*

b. Bagaimana pembuatan ontologi pada domain produk *skincare* untuk membantu pengguna dalam pencarian produk *skincare* sesuai kebutuhan yang terdapat pada Female Daily*?*

1.3 Batasan Masalah

Berdasarkan identifikasi rumusan masalah, penulis memberikan batasan masalah dari penelitian sebagai berikut:

a. Penelitian ini akan membatasi penggunaan data dari komentar produk *skincare* yang mendapat penghargaan dalam *Best of Beauty Awards* di Female Daily dengan periode tahun 2020 hingga 2024.

b. Metode yang digunakan untuk *topic modeling* yaitu metode *Latent Dirichlet Allocation* (LDA).

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, tujuan dari penelitian ini yaitu:

a. Mengidentifikasi topik-topik utama dalam komentar produk *skincare* yang mendapat penghargaan dalam *Best of Beauty Awards* di Female Daily dengan tujuan untuk menggali topik diskusi yang mencerminkan pengalaman pengguna.
b. Membuat ontologi pada domain produk *skincare* untuk membantu pengguna dalam pencarian produk *skincare* yang terdapat pada Female Daily.

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

a. Memberikan wawasan kepada pembaca mengenai *topic modeling* terkait komentar produk *skincare* pemenang *Best of Beauty Awards* di Female Daily, terutama dalam mengidentifikasi tema-tema utama seperti efektivitas produk, kecocokan dengan jenis kulit, dan pengalaman penggunaan.
b. Menyusun ontologi pada domain produk *skincare* berdasarkan hasil *topic modeling* untuk memetakan konsep, relasi, dan kata kunci yang sering digunakan oleh pengguna dalam komentar di Female Daily.

1.6 Sistematika Penulisan

Penulisan penelitian ini disusun dari beberapa bagian yang terbagi dalam 5 bab:

BAB I PENDAHULUAN

Bagian pendahuluan berisi latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

BAB II LANDASAN TEORI

Bagian landasan teori berisi penjelasan konsep dasar yang digunakan untuk merumuskan solusi terhadap permasalahan penelitian dan tinjauan terhadap sejumlah penelitian terdahulu yang memiliki relevansi atau kemiripan topik dengan studi ini.

BAB III METODOLOGI PENELITIAN

Bagian metodologi penelitian berisi langkah-langkah yang dilakukan peneliti seperti pengambilan data, *preprocessing*, TF-IDF, pemodelan topik menggunakan LDA, analisis hasil, interpretasi topik, dan pembuatan ontologi.

BAB IV HASIL DAN PEMBAHASAN

Bagian hasil dan pembahasan berisi penjelasan mengenai implementasi metode dan hasilnya.

BAB V KESIMPULAN DAN SARAN

Bagian ini merupakan bab terakhir yang membahas kesimpulan dan saran terhadap penelitian yang telah dilakukan pada tugas akhir.

BAB II
LANDASAN TEORI

2.1 Female Daily

Dilansir dari laman *femaledaily.com/about* mengenai penjelasan perjalanan awal dibentuk, Female Daily bermula dari sebuah blog pribadi berisi konten *fashion* dan kecantikan yang dikelola oleh Hanifa Hambadar pada tahun 2005. Melalui platform *multiplay.com*, Hanifa bertemu dengan blog milik Affi Assegaf yang juga seorang *blogger* dan sedang bekerja di MAC *Cosmetics*. Pada tahun 2007, Hanifa dan Affi bekerjasama dan mendirikan blog Female Daily. Melihat potensi dari blog tersebut yang dapat menjadi sebuah bisnis karena banyaknya pengiklan yang berminat menaruh produk mereka pada blog, akhirnya Hanifa dan Affi mengembangkan blog mereka dan membuka kantor, membuat PT, dan mengangkat orang sebagai karyawan untuk bagian penjualan. Setelah itu mereka meluncurkan blog bernama MommiesDaily yang anggotanya masih menggunakan fasilitas forum FashioneseDaily untuk berdiskusi seputar kehamilan dan lain-lain. Melihat banyak topik di forum yang melebar, akhirnya Hanifa dan Affi memutuskan untuk memindahkan domain ke *femaledaily,com* (Harits, 2016).

Gambar 2.1.1 Logo Female Daily

Pada tahun 2015, Female Daily hanya memfokuskan pada bidang kecantikan dan mengembangkan blog menjadi sebuah aplikasi bernama ‘Female Daily - *Beauty Review*’ yang tersedia di perangkat berbasis Android dan iOS. Aplikasi ini telah mencapai jumlah 1 juta unduhan di tahun 2023 dan jumlah anggota pengguna yang terdaftar lebih dari 990 ribu. Female Daily menyediakan lebih dari 2,5 juta ulasan produk dari para anggotanya. Hingga saat ini, tercatat lebih dari 100 ribu produk telah diulas, dengan kategori *skincare* sebagai yang paling dominan, yaitu lebih dari 1,35 juta ulasan (Female Daily, t.t.). Hal ini menunjukkan pertumbuhan yang signifikan dan menunjukkan bahwa Female Daily telah berhasil membangun sebuah komunitas yang kuat dan loyal. Bagi para pengunjung laman

femaledaily.com atau pengguna *smartphone* yang telah memiliki aplikasi Female Daily, diharuskan bergabung terlebih dahulu menjadi anggota agar dapat berkontribusi memberikan *review* (ulasan) pada suatu produk. Selain itu, Female Daily ini juga menawarkan fitur-fitur utama seperti:

a. Platform Beauty Reviews (FD *Review*)

   Pada platform FD *Review* atau dapat diakses pada laman reviews.femaledaily.com, pengguna dapat mencari informasi mengenai detail produk yang ingin dibeli dan mendapat informasi tentang ulasan, tipe kulit, jangka pemakaian, dan penilaian bintang dari pengguna lain mengenai sebuah produk yang telah mereka gunakan.

b. Konten editorial (FD Editorial)

   Pada platform FD Editorial atau dapat diakses pada laman editorial.femaledaily.com, berisi laman yang menyajikan informasi seputar kecantikan yang ditulis oleh tim editor Female Daily.

c. *E-Commerce* (FD Studio)

   Pada platform FD Studio atau dapat diakses pada laman studio.femaledaily.com, laman ini menyediakan jual beli berbagai produk kecantikan.

d. Forum diskusi (FD *Talk*)

   Pada platform FD Talk atau dapat diakses pada laman talk.femaledaily.com, komunitas *online* yang berisikan berbagai *thread* atau utas mengenai topik seputar perempuan dan kecantikan.

Female Daily rutin menyelenggarakan survei setiap tahunnya yang hasilnya disajikan dalam ajang Female Daily *Best of Beauty Awards* (BoBA). Acara ini bertujuan untuk mengevaluasi produk kecantikan yang paling banyak dipilih lewat pemungutan suara atau voting dan penilaian juri. Para *member* Female Daily dapat memilih produk terbaik yang pernah digunakan, atau dapat juga memasukkan produk tertentu yang pernah digunakan namun tidak ada di pilihan. Pemilik merek juga dapat mengajukan produk-produknya untuk menjadi nominasi dengan syarat dan ketentuan yang berlaku. Dengan demikian, Female Daily *Best of Beauty Awards* tidak hanya menjadi barometer kualitas produk, tetapi juga bentuk apresiasi terhadap kerja keras, inovasi, dan keunggulan yang ditawarkan oleh berbagai merek, oleh individu, dan komunitas dengan maksud memajukan industri kecantikan. Melalui penghargaan ini, Female Daily berupaya mendorong peningkatan kualitas produk kecantikan di pasaran.

2.2 *Electronic Word of Mouth* (e-WOM)

Menurut Kotler dan Amstrong dalam Erdawati dkk. (2023), *electronic word of mouth* (e-WOM) merupakan bentuk komunikasi pemasaran dari mulut ke mulut yang disampaikan melalui media digital, seperti situs web, iklan daring, aplikasi seluler, video, email, blog, media sosial, serta kegiatan promosi lainnya yang bersifat menarik sehingga mendorong konsumen untuk membagikannya kepada orang lain. *Electronic word of mouth* memiliki pengaruh yang kuat dan kredibel terhadap keputusan pembelian konsumen karena mampu menyebarkan informasi mengenai kualitas produk dan layanan secara cepat, efektif, serta dengan biaya yang relatif rendah (Wahyu & Putri, 2023). Informasi yang disampaikan melalui *electronic word of mouth* (eWOM) berperan sebagai referensi penting bagi konsumen dalam mengevaluasi produk atau merek, karena tidak hanya memberikan gambaran mengenai kualitas barang atau jasa, tetapi juga secara efektif membantu mengurangi risiko dan ketidakpastian yang dihadapi sebelum konsumen memutuskan untuk melakukan pembelian (Sualang dkk., 2023).

2.3 Text Mining

Text mining dapat diartikan sebagai suatu proses mengekstraksi informasi yang bernilai dari sekumpulan dokumen teks melalui identifikasi dan eksplorasi pola-pola yang menarik dengan memanfaatkan berbagai teknik analisis (Rosalinda dkk., 2023). *Text mining* melibatkan algoritma *data mining*, NLP, *machine learning*, dan operasi statistik untuk mengekstrak konten yang bernilai dari format tidak terstruktur seperti data teks. Oleh karena itu, *text mining* dapat meningkatkan tren dan aktivitas komersial dengan mengekstrak informasi dari konten yang dihasilkan pengguna atau *user-generated content* (UGC) (Albalawi dkk., 2020). *Text mining* meliputi tahapan penataan teks *input*, pencarian pola dalam data teks yang telah terstruktur, serta tahap akhir berupa evaluasi dan interpretasi hasil. Tujuan utama dari *text mining* yaitu mengekstrak informasi bernilai dari teks yang kemudian diubah menjadi pengetahuan yang dapat dimanfaatkan dalam proses pengambilan keputusan.

2.4 Topic Modeling

Topic Modeling merupakan sebuah metode yang menggunakan berbagai algoritma untuk mengidentifikasi dan menandai tema-tema utama yang tersembunyi di dalam kumpulan teks (Kherwa & Bansal, 2020). Algoritma *topic modeling* menawarkan beragam cara untuk menggali informasi yang tidak langsung tampak dalam kumpulan teks. Dengan algoritma ini, dapat diidentifikasi dan dikelompokkan tema-tema utama yang mendasari dokumen-dokumen

tersebut. Metode pemodelan topik dapat menemukan tema-tema utama dalam kumpulan dokumen tanpa perlu terlebih dahulu memberi label atau keterangan pada dokumen tersebut, sehingga metode ini dapat digunakan untuk mengorganisasi dan meringkas arsip elektronik dalam jumlah besar yang tidak mungkin dilakukan oleh manusia (Blei, 2012). *Topic modeling* melihat setiap dokumen sebagai campuran dari beberapa topik dan kata (Puspita dkk., 2024). Implementasi *topic modeling* bergantung pada pemilihan metode yang tepat. Penggunaan metode yang sesuai akan secara signifikan meningkatkan kualitas hasil pemodelan topik, sehingga dapat menjadi dasar yang lebih kuat untuk analisis maupun pengambilan keputusan yang berbasis data. Menurut Kherwa & Bansal (2020), metode pemodelan topik yang paling populer dan berkontribusi pada setiap aspek analisis teks di berbagai domain sebagai berikut.

a. *Latent Semantic Analysis* (LSA)

   Metode LSA merupakan suatu metode aljabar yang didasarkan pada *single value decomposition* (SVD) dan menyajikan ruang semantik dan penggunaan kontekstual kata-kata cenderung lebih dekat satu sama lain (Kherwa & Bansal, 2020). Berbeda dengan manusia yang belajar bahasa secara alami, mesin perlu mempelajari semua makna dan konteks bahasa dari data. Prinsip metode ini yaitu menghitung jarak antara vektor-vektor untuk mengelompokkan dokumen berdasarkan topik yang relevan (Novarian dkk., 2023). Cara kerja metode LSA dengan cara mengurangi kebisingan (*noise*) pada kumpulan data yang disajikan dalam bentuk matriks melalui penggunaan *Singular Value Decomposition* (SVD) (Jayadianti dkk., 2020). LSA bertujuan untuk mengidentifikasi ide atau topik tersembunyi dalam suatu dokumen dengan cara memanfaatkan konteks kata-kata di dalamnya.

b. *Non Negative Matrix Factorization* (NMF)

   Metode NMF merupakan suatu metode *unsupervised matrix factorization* (berbasis aljabar linear) yang mampu melakukan reduksi dimensi dan pengelompokan secara bersamaan (Albalawi dkk., 2020). Albalawi juga mengatakan model ini dirancang untuk mengidentifikasi tema-tema utama dalam kumpulan data teks yang singkat. NMF bekerja dengan cara memecah data menjadi tiga matriks utama, yaitu matriks korelasi istilah yang memperhitungkan hubungan asimetris antar istilah, matriks istilah (dokumen yang menunjukkan frekuensi kemunculan istilah dalam setiap dokumen), serta matriks *bag-of-words* yang merepresentasikan seluruh kumpulan dokumen sebagai kumpulan kata-kata. Dengan menggunakan ketiga matriks ini, model NMF dapat mengungkap struktur laten dari data teks dan mengidentifikasi topik-topik yang mendasarinya. NMF merupakan

representasi matematis yang menghubungkan kata-kata dengan topik-topik tertentu, di mana semua nilai dalam matriks ini bersifat nonnegatif. Sifat nonnegatif ini memungkinkan untuk menginterpretasikan hubungan antara kata dan topik secara lebih intuitif dan mudah dipahami (Patricia, 2021).

c. *Probabilistic Latent Semantic Analysis* (PLSA)

Metode PLSA merupakan pengembangan lebih lanjut dari *Latent Semantic Analysis* (LSA) yang menggunakan pendekatan probabilistik (Figuera & Bringas, 2024). LSA awalnya menggunakan metode hitung kata-kata yang muncul bersamaan dalam dokumen-dokumen untuk membentuk matriks frekuensi. Selanjutnya, matriks ini diuraikan menggunakan *Singular Value Decomposition* (SVD) untuk mendapatkan dimensi-dimensi laten yang dapat diinterpretasikan sebagai aspek-aspek semantik. PLSA mengambil konsep dasar ini dan memperluasnya dengan menggunakan model probabilistik untuk memberikan pemahaman yang lebih baik tentang hubungan antara kata-kata dan dokumen. Menurut Figuera & Bringas (2024) kemampuan beradaptasi yang tinggi, kejelasan hasil, serta sifat statistik yang kokoh dari PLSA telah membuka peluang luas untuk penerapan model ini dalam berbagai domain. Dalam penerapannya, PLSA memungkinkan kita untuk memetakan konsep kata dan dokumen ke dalam entitas diskrit lainnya, sehingga kita dapat menguji dan mengkonfirmasi kebenaran asumsi-asumsi yang mendasari model PLSA.

d. *Latent Dirichlet Allocation* (LDA)

Menurut Wahyudi dkk. (2021), LDA merupakan sebuah model probabilistik generatif yang secara generatif dapat menghasilkan kumpulan tulisan yang disebut korpus. Prinsip LDA mengelompokkan dokumen ke dalam kategori berdasarkan tema atau topik yang relevan (Novarian dkk., 2023). *Latent* merujuk pada segala sesuatu informasi yang tersembunyi dalam data, *Dirichlet* merujuk pada distribusi topik dalam dokumen dan distribusi kata dalam topik, dan *Allocation* mengacu pada mengalokasikan topik ke dokumen dan kata-kata dalam dokumen ke topik. Metode LDA dapat menghasilkan sekumpulan topik yang mendeskripsikan keseluruhan korpus, di mana masing-masing topik dapat dipahami secara individual dan juga mampu menangani korpus dokumen-kata berskala besar tanpa memerlukan pelabelan teks apa pun (Albalawi dkk., 2020). Dikutip oleh Puspita dkk. (2024), model LDA dinilai lebih unggul dalam menghasilkan topik dengan makna yang logis, kemampuan yang lebih baik dalam menafsirkan hasil analisis, dan memprediksi pola data di masa depan.

2.5 Preprocessing

Preprocessing merupakan suatu proses awal sebelum analisis data dilakukan dengan mengorganisasikan data yang akan diolah. *Preprocessing* diterapkan untuk membersihkan data mentah yang tidak sempurna, tidak konsisten, atau bermasalah sebelum informasi yang relevan diekstrak dan dikonversi ke dalam format terstruktur untuk analisis lebih lanjut. Analisis terhadap data terstruktur ini bertujuan untuk mengungkap pola-pola yang tampak maupun tersembunyi di dalam data (Albalawi dkk., 2020). *Preprocessing* terdiri dari dari beberapa tahapan, di antaranya *remove punctuation, case folding, tokenizing, stemming,* dan *stopwords.*

a. Remove Punctuation

Tahap ini bertujuan untuk menghilangkan data dari elemen-elemen non-alfanumerik seperti angka, simbol khusus, tanda baca, dan *hyperlink* dari teks, sehingga menghasilkan korpus teks yang lebih murni dan konsisten yang hanya terdiri dari karakter alfabet (Novarian dkk., 2023).

b. Case Folding

Tahap *case folding* berfungsi untuk mengubah seluruh karakter kata menjadi huruf kecil. Tujuannya untuk menghindari kesalahan dalam pengenalan kata karena perbedaan kapitalisasi, seperti pada kata ‘Data’ dan ‘data’ yang pada dasarnya merujuk pada entitas yang sama (Chilmi, 2021).

c. Tokenizing

Tahapan *tokenizing* melakukan pemenggalan teks input menjadi satuan kata-kata individu yang disebut “token”. *Tokenizing* bertujuan untuk memudahkan saat proses perhitungan kata hingga transformasi data (Chilmi, 2021).

d. Normalization

Tahap *normalization* merupakan tahapan untuk mengubah kata-kata yang tidak normal menjadi sesuai dengan standar (Puspita dkk., 2024). Data komentar umumnya mengandung berbagai bentuk kata tidak standar seperti singkatan, slang, kesalahan pengetikan (*typo*), dan pengulangan karakter, misalnya kata "udh" menjadi "sudah", "murce" menjadi "murah", "kurg" menjadi "kurang", dan "kereeennn" menjadi "keren".

Proses ini penting untuk meminimalkan variasi bentuk kata yang semakna sehingga mendukung akurasi hasil dalam tahap pemrosesan berikutnya.

e. Stopwords

Tahap *stopwords* merupakan proses eliminasi kata-kata fungsional yang memiliki frekuensi kemunculan tinggi namun kontribusi rendah terhadap makna semantik suatu teks (Puspita dkk., 2024). Dalam tahap ini, dibuat *stoplist* yang berisikan daftar kata-kata yang tidak relevan atau tidak penting. Tujuan tahap ini untuk memproses kata-kata tidak penting dan kata-kata yang tersisa akan dianggap sebagai kata-kata penting atau *keywords* (Chilmi, 2021).

f. Stemming

Pada tahap *stemming* dilakukan pembersihan data dari kata berimbuhan, awalan, sapaan, akhiran, atau kombinasi. Sehingga setiap kata dalam dokumen hanya berupa kata dasarnya (Yutika dkk., 2021).

2.6 *Term Frequency-Inverse Document Frequency* (TF-IDF)

Metode *Term Frequency-Inverse Document Frequency* atau TF-IDF merupakan suatu metrik numerik yang digunakan untuk mengukur bobot atau pentingnya suatu kata (*term*) dalam suatu konten dari kumpulan dokumen. Pengukuran bobot *term* pada sebuah dokumen mencerminkan tingkat kepentingan suatu kata terhadap dokumen, semakin sering suatu kata muncul maka semakin tinggi bobot yang diberikan, sehingga mengindikasikan bahwa kata tersebut memiliki kontribusi yang besar terhadap makna dokumen. Metrik ini mempertimbangkan dua faktor utama yaitu *Term Frequency* (TF) atau frekuensi kemunculan kata dalam suatu dokumen dan *Inverse Document Frequency* (IDF) atau frekuensi dokumen yang mengandung kata tersebut (Albalawi dkk., 2020). Bobot *term i* dihasilkan dengan cara mengalikan nilai TF dengan IDF menggunakan persamaan (1) (Supono & Suprayogi, 2021).

$$WW_{ii,jj} = ttt_{i,jj} \times iiiit_{ii} \quad (1)$$

Keterangan:

$WW_{ii,jj}$ = Bobot *term i* pada dokumen *j*

$ttt_{i,jj}$ = Jumlah kemunculan *term i* pada dokumen *j*

$iiiitt_{ii}$ = *Inverse* $iitt_{ii}$

2.6.1 *Term Frequency* (TF)

Tahap pertama dalam TF-IDF yaitu *Term Frequency* (TF) atau frekuensi kemunculan kata dalam yang muncul dalam suatu dokumen (Fahrezi dkk., 2024). Rumus dari TF ditujukkan pada persamaan (2).

$$ttt_{ii,jj} = \frac{jjjjjjjjjjh\ kkkkjjjjkkkkjjjjjjkk\ ttkkttjj\ ii\ iijjjjjjjjj\ iiddkkjjjjkkkk\ ii}{ttddttjjjj\ jjjjjjjjjh\ ttkkttjj\ iijjjjjjjj\ iiddkkjjjjkkkk} \quad (2)$$

2.6.2 *Inverse Documnet Frequency* (IDF)

Tahap selanjutnya yaitu *Inverse Document Frequency* (IDF) atau frekuensi dokumen yang mengandung kata tersebut. IDF merupakan suatu metode statistika numerik yang menghitung seberapa diperlukannya kata dalam sebuah dokumen (Supono & Suprayogi, 2021). Rumus dari IDF ditujukkan pada persamaan (3).

$$iiiitt_{tt} = jjddll\left(\frac{NN}{iitt_{ii}}\right) \quad (3)$$

Keterangan:

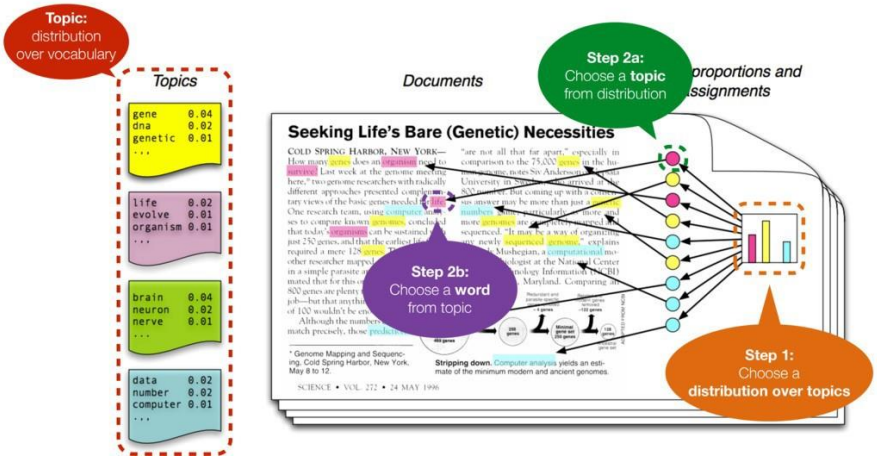
$iiiitt_{tt}$ = *Inverse Document Frequency*

N = Jumlah Dokumen

$iitt_{ii}$ = Banyaknya dokumen yang memuat *term* ke-*i*

2.7 *Latent Dirichlet Allocation* (LDA)

Latent Dirichlet Allocation atau LDA merupakan model probabilistik yang pertama kali diusulkan oleh Blei dkk. (2003). LDA dianggap sebagai algoritma pemodelan topik paling populer untuk mengekstrak topik dari kumpulan dokumen karena memberikan hasil yang akurat dan dapat dilatih secara *online*. LDA merupakan algoritma statistik yang bersifat generatif dan *unsupervised* yang digunakan untuk mengungkap tema-tema laten dari kumpulan dokumen dengan pendekatan *Bayesian* (Albalawi dkk., 2020). LDA memanfaatkan distribusi Dirichlet untuk menghasilkan distribusi topik pada setiap dokumen. Proses generatif yang dihasilkan oleh distribusi Dirichlet kemudian digunakan untuk mendistribusikan kata-kata dalam suatu dokumen ke berbagai topik yang berbeda. Objek yang mampu diamati oleh LDA adalah dokumen, sedangkan distribusi kata dengan topik, distribusi topik dengan dokumen, serta penentuan topik dari setiap kata bersifat tidak terlihat atau tersembunyi (*latent*), maka model ini dinamakan *Latent Dirichlet Allocation* (Blei, 2012).

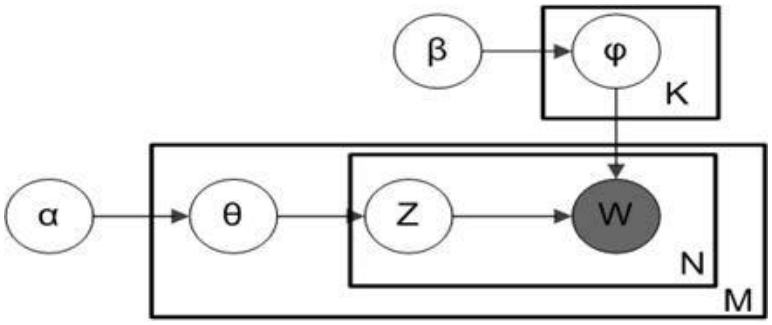


Gambar 2.2 Ilustrasi LDA
Sumber: Blei (2012)

Model LDA mengasumsikan bahwa setiap dokumen terdiri dari sejumlah topik yang berbeda, di mana setiap topik merupakan distribusi probabilitas atas kata-kata seperti yang terlihat pada Gambar 2.2. Untuk setiap dokumen di dalam koleksi, akan menghasilkan kata-kata dalam dua proses tahapan:

1. Secara acak memilih distribusi atas topik-topik.
2. Untuk setiap kata di dalam dokumen:
   a. Secara acak dipilih sebuah topik dari distribusi atas topik pada langkah 1.
   b. Secara acak dipilih distribusi sebuah kata dari distribusi yang sesuai atas kosakata.

Visualisasi pada Gambar 2.2, model LDA dapat digambarkan sebagai sebuah grafik Bayesian berlapis tiga, di mana setiap variabel acak dalam model diwakili oleh sebuah node, dan hubungan probabilistik antara variabel-variabel tersebut dihubungkan oleh tepi. Struktur grafik ini menunjukkan bahwa parameter α dan β hanya diambil sampel sekali untuk seluruh korpus dokumen, sedangkan topik untuk setiap dokumen dan kata-kata dalam setiap dokumen diambil sampel secara berulang berdasarkan topik yang telah ditentukan sebelumnya (Setijohatmo dkk., 2020).



Gambar 2.3 Visualisasi Grafik Model LDA
Sumber: Setijohatmo dkk. (2020)

LDA digambarkan dengan model grafis menggunakan *plate notation* seperti pada Gambar 2.3.

Keterangan:

$\beta\beta$ = *Dirichlet parameter* atas distribusi kata terhadap topik

$\varphi\varphi$ = Distribusi kata terhadap topik dalam *corpus*

K = Kumpulan topik

W = Kata

N = Kumpulan kata

M = Kumpulan dokumen

Z = Topik *index assignment*

$\theta\theta$ = Dokumen

jj = *Dirichlet parameter* atas distribusi topik terhadap dokumen

Latent Dirichlet Allocation atau LDA dapat dirumuskan sebagaimana ditujukkan pada persamaan (4) (Rosalinda dkk., 2023).

$$pp(ww, zz, \theta\theta, \varphi\varphi|\alpha\alpha, \beta\beta) = pp(\theta\theta|\alpha\alpha)\, pp(zz|\theta\theta)\, pp(\varphi\varphi|\beta\beta)\, pp(ww|zz, \varphi\varphi) \quad (4)$$

Persamaan (4) memperlihatkan bahwa terdapat empat komponen pada ruas sebelah kanan persamaan tersebut. Pertama, distribusi topik terhadap dokumen ($\theta\theta$) dibangkitkan mengikuti distribusi *Dirichlet* dengan parameter $\alpha\alpha$. Nilai probabilitasnya dihitung dengan rumus pada persamaan (5).

$$pp(\theta\theta|\alpha\alpha) = \frac{\Gamma(\sum_{k}^{KK} = 1\ \alpha\alpha_{kk})}{\prod_{kk}^{KK} = 1\ (\alpha\alpha_{kk})} \square \ \underset{kk}{\overset{KK}{=}}\ \theta\theta_{kk}^{\alpha\alpha_{kk}-1}, \alpha\alpha_{kk} > 0 \quad (5)$$

Kedua, distribusi probabilitas suatu kata terhadap dokumen untuk masing-masing topik (zz) bergantung pada distribusi $\theta\theta$ tersebut di atas. Setiap kata ww_{nn} dalam dokumen dari N kata diberi nilai dari 1,…,K. Distribusi peluang untuk zz yang diamati dengan $\theta\theta$ yang diamati mengikuti distribusi Multinomial yang dihitung menggunakan persamaan (6).

$$pp(zz|\theta\theta) = \prod_{mm=1}^{MM} \prod_{kk=1}^{KK} \theta\theta_{mm,kk}^{nn_{mm,kk}} \quad (6)$$

Keterangan:

$\theta\theta_{mm,kk}$ = distribusi topik k terhadap dokumen m.

$kk_{mm,kk}$ = jumlah kemunculan topik k yang ditentukan berdasarkan kata-kata dalam dokumen m.

Ketiga, distribusi kata terhadap topik ($\varphi\varphi$) juga dibangkitkan mengikuti distribusi *Dirichlet* dengan parameter $\beta\beta$. Probabilitas $\varphi\varphi$ untuk seluruh topik dan seluruh kata dalam kosakata diformulasikan pada persamaan (7).

$$pp(\varphi\varphi|\beta\beta) = \prod_{kk}^{KK} = 1 \frac{\Gamma(\beta\beta_{kk})}{\prod_{vv}^{VV} = 1 \Gamma(\beta\beta_{kk,vv})} \prod_{vv=1}^{VV} \varphi\varphi_{kk,vv}^{\beta\beta_{kk,vv}-1} \quad (7)$$

Keterangan:

$\varphi\varphi_{kk,vv}$ = probabilitas *term* vv terhadap topik k dan $\beta\beta_{kk}>0$.

Terakhir, distribusi peluang kata-kata (w) dalam korpus terhadap topik tertentu (zz)bergantung pada distribusi kata terhadap topik ($\varphi\varphi$). Distribusi tersebut mengikuti distribusi Multinomial yang didapatkan menggunakan persamaan (8).

$$pp(ww|zz, \varphi\varphi) = \prod_{kk=1}^{KK} \prod_{vv=1}^{VV} \varphi\varphi_{kk,vv}^{nn_{kk,vv}} \quad (8)$$

Keterangan:

$\varphi\varphi_{kk,vv}$ = probabilitas *term* v terhadap topik k

2.8 Web Semantik

Web semantik merupakan perluasan dari *World Wide Web* konvensional yang bertujuan untuk memberikan makna yang lebih jelas dan terstruktur pada informasi yang ada di dalamnya. Perkembangan pesat internet telah menghasilkan volume data yang sangat besar, namun sebagian besar data ini disajikan dalam format yang mudah dibaca oleh manusia (*human-readable*) dan belum tentu dapat dipahami oleh komputer. Untuk mengatasi permasalahan ini, teknologi web semantik hadir sebagai solusi dengan menyediakan kerangka kerja untuk memberikan makna yang jelas pada data. Dengan demikian, baik manusia maupun

komputer dapat lebih mudah memahami dan menggunakan informasi tersebut. Konsep utama dari web semantik yaitu membuat data di internet menjadi lebih terdefinisi saling terhubung, sehingga komputer dapat memproses, mengintegrasikan, dan memanfaatkan data ini secara otomatis untuk berbagai keperluan (Pramartha, 2020).

2.9 Ontologi

Ontologi berperan sebagai kerangka kerja fundamental dalam web semantik, berfungsi sebagai penghubung antara konsep-konsep yang dipahami manusia dengan representasi data yang dapat diproses oleh komputer. Dengan kata lain, ontologi bertindak sebagai jembatan antara pemahaman manusia dan kemampuan pemrosesan mesin (Pramartha, 2020). Selain itu, penggunaan ontologi dapat meningkatkan akurasi hasil pencarian informasi di internet.

Sebuah konsep dalam ontologi dapat memiliki objek atau *instance.* Secara teknis, ontologi dimodelkan menggunakan konstruksi seperti *class, property, facet,* dan *instance. Class* menunjukkan konsep atau makna dari suatu domain tertentu. *Class* merupakan sekumpulan elemen yang memiliki sifat yang sama. Sebuah *class* dapat memiliki *sub class* yang menunjukkan konsep lebih spesifik. *Property* merepresentasikan relasi antara dua individu yang menghubungkan suatu individu dari domain tertentu dengan individu dari *range* tertentu. *Facet* merepresentasikan batasan mengenai *property*. *Instance* merupakan objek dari suatu kelas (Cahyono dkk., 2016).

Dalam pengaplikasiannya, model memerlukan sebuah *human-readable medium* untuk memudahkan pengguna dalam memahami informasi yang diberikan dari model ontologi. Sebelum informasi tersebut diberikan, perlu dilakukan evaluasi terhadap kebenaran informasi tersebut. Evaluasi ini umumnya menggunakan dua *tools* seperti Protégé dan aplikasi yang dibuat, misalnya aplikasi berbasis web. Hal ini untuk memastikan bahwa informasi yang disampaikan akurat dan relevan dengan model ontologi yang telah dibangun (Rachman & Nurjanah, 2019).

2.10 *Ontology Web Language* (OWL)

Ontology Web Language (OWL) merupakan bahasa yang populer digunakan ketika membuat ontologi. Tujuan OWL hampir sama dengan RDF *Schema* yaitu untuk mendefinisikan ontologi termasuk *class*, *property*, dan hubungan untuk setiap aplikasi yang spesifik. OWL dapat menyatakan relasi yang lebih kompleks (Muryanti dkk., 2021).

2.11 SPARQL

SPARQL merupakan bahasa *query* yang direkomendasikan oleh W3C (*World Wide Web Consortium*) sebagai standar untuk mengakses dan mengekstraksi informasi dari data berbasis graph RDF pada Semantic Web. Dengan menggunakan SPARQL, pengguna dapat mengambil data dari sumber yang terstruktur maupun semi-terstruktur. Hasil query yang dieksekusi dapat dikembalikan dalam berbagai format keluaran seperti XML, JSON, RDF, atau HTML (Himawan dkk., 2020).

2.12 Protégé

Protégé merupakan sebuah perangkat lunak yang dirancang untuk membangun dan memvisualisasikan ontologi serta taksonomi berbasis web, dikembangkan oleh Stanford's Biomedical Informatics Research (BMIR). Protégé digunakan untuk memodelkan ontologi menggunakan editor Protegé-Frames dan Protegé-OWL. Ontologi yang telah dibangun kemudian dapat diekspor dalam berbagai format standar seperti RDF, RDFS, OWL, dan skema XML, sehingga dapat diintegrasikan dengan sistem informasi lainnya (Santosa dkk., 2023).

2.13 Python

Bahasa pemrograman Python pertama kali dikembangkan oleh Guido van Rossum di Pusat Matematika dan Komputer (CWI), Amsterdam pada tahun 1991 (Awangga dkk., 2019). Python merupakan bahasa pemrograman yang memanfaatkan interpreter untuk mengeksekusi instruksi. Interpreter Python memiliki kemampuan untuk menerjemahkan kode secara langsung yang dapat dieksekusi oleh komputer. Fleksibilitas Python menjadikan bahasa ini dapat dijalankan di berbagai platform seperti Linux, Windows, dan lainnya. Python bersifat *open source* yang menjadi salah satu keunggulannya, karena dalam pengembangannya melibatkan para *programmer*, peneliti, dan pengguna dari berbagai latar belakang.

Python memiliki kemampuan yang sangat baik dalam berinteraksi dengan berbagai jenis *database* dan dapat dengan mudah memanipulasi berbagai format *file*. Hal ini menjadikan Python sebagai pilihan yang populer untuk mengembangkan prototipe atau perangkat lunak secara cepat dan andal. Selain itu, Python sangat digemari para peneliti berkat kemampuannya dalam mengolah data dalam jumlah besar dan melakukan perhitungan matematis yang rumit (Rahman dkk., 2023).

2.14 Penelitian Terdahulu

Dalam kajian ini penulis membahas sejumlah studi terdahulu sebagai dasar perbandingan dan pijakan dalam penyusunan metode penelitian yang digunakan. Kajian ini juga membahas kelebihan dan kekurangan dari metode yang digunakan oleh penelitian terdahulu. Proses penelusuran literatur dilakukan melalui Google Scholar, ResearchGate, IEEE Xplore, dan Elsevier dengan rentang publikasi tahun 2019 hingga 2025 guna memastikan bahwa sumber yang digunakan bersifat relevan dan terkini. Kata kunci yang digunakan dalam proses pencarian yaitu *topic modeling*, ulasan pengguna, kosmetik, *skincare*, dan ontologi. Pemilihan kata kunci tersebut disesuaikan dengan ruang lingkup penelitian, yaitu analisis topik dari ulasan produk *skincare* serta pembuatan ontologi. Hasil pencarian kemudian diseleksi berdasarkan kesesuaian tema, pendekatan metodologis, dan kontribusi yang dapat dibandingkan atau dijadikan acuan dalam penelitian ini. Literatur terdahulu yang telah dikumpulkan disajikan dan dijelaskan pada Tabel 2.1.

Tabel 2.1 Bingkai Analisis Berdasarkan Metode *Topic Modeling* (Bagian 1)

Sumber	LDA	PLSA	LSA	NMF	Hasil
Puspita dkk. (2024)	✓				Peneliti menggunakan metode *Latent Dirichlet Allocation* (LDA) dan memanfaatkan nilai *coherence* untuk menganalisis dan mengevaluasi jumlah topik terbaik. 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. Dapat disimpulkan bahwa jumlah topik 6 menghasilkan topik-topik yang relevan mengenai merek Somethinc.
Destarani dkk. (2019)	✓				Penelitian ini menggunakan model LDA untuk mengidentifikasi topik utama dalam keluhan warga Denpasar. Hasilnya menunjukkan bahwa topik tren berada pada topik 4 dengan nilai probabilitas topik sebesar 0,055. Terdapat 78,95% dari topik yang dihasilkan dapat diinterpretasikan dengan baik sesuai dengan data keluhan di situs Pro Denpasar dan terdapat 21,05% topik yang tidak dapat diinterpretasikan dengan benar.

Tabel 2.1 Bingkai Analisis Berdasarkan Metode *Topic Modeling* (Bagian 2)

Sumber	LDA	PLSA	LSA	NMF	Hasil
Tondang dkk. (2023)	✓				Penelitian ini mengevaluasi *topic modeling* pada ulasan aplikasi BNI *Mobile Banking*, BCA Mobile, dan BRImo dengan menggunakan LDA (*Latent Dirichlet Allocation*). Metode LDA digunakan untuk mengelompokkan ulasan ke dalam topik-topik yang mencerminkan pengalaman pengguna dengan aplikasi *mobile banking*. Analisis data menunjukkan adanya perbedaan dan kesamaan dalam kualitas layanan antara ketiga aplikasi tersebut.
Mustahidah (2021)	✓	✓			Penelitian ini bertujuan untuk mendapatkan hasil sentimen dari ulasan tamu hotel dari situs Tripadvisor, yaitu Gino Feruci Braga, Golden Flower, dan Belviu Hotel menggunakan metode NBC serta mendapatkan hasil topik dalam data ulasan positif dan data ulasan negatif pada masing-masing hotel menggunakan LDA dan PLSA. Berdasarkan hasil indeks *Calinski-Harabasz* dapat disimpulkan bahwa secara konsisten, LDA memperoleh nilai indeks yang lebih tinggi dibandingkan PLSA pada setiap hotel, baik untuk ulasan positif maupun negatif.
Mohammed & Al-Augby (2020)	✓		✓		Penelitian ini membandingkan metode klasifikasi dokumen teks ilmiah tidak terstruktur (*e-books*) berdasarkan teks lengkapnya dengan menerapkan pendekatan pemodelan topik LDA dan LSA. Dataset pada penelitian ini terdiri dari 100 buku yang berisi sekitar 1 juta kata berdasarkan teks lengkap. Hasil penelitian menunjukkan bahwa LDA memiliki hasil yang lebih baik dibandingkan LSA. Nilai koherensi terbaik dengan jumlah topik 20 yang diperoleh dari metode LDA adalah 0,54846, sedangkan nilai koherensi LSA adalah 0,4047.
Nugumanova dkk. (2022)				✓	Penelitian ini menerapkan lima algoritma NMF pada *dataset* ACTER yang berisi dokumen khusus bidang dengan anotasi terminologi. Tujuannya adalah mengembangkan metode *unsupervised* untuk mengekstraksi istilah tanpa anotasi manual. Dengan menguji variasi jumlah topik dan kata per topik, hasil terbaik NMF meraih F1-score 0,47, mendekati performa dua metode *supervised* yang lebih kompleks. Temuan ini menunjukkan bahwa NMF efektif dalam mengekstraksi istilah relevan dengan pengawasan minimal.

Tabel 2.1 Bingkai Analisis Berdasarkan Metode *Topic Modeling* (Bagian 3)

Sumber	LDA	PLSA	LSA	NMF	Hasil
Wu dkk. (2020)				✓	Penelitian lain menerapkan model *robust semi-supervised NMF* (rssNMF) untuk mengelompokkan data *single-cell RNA-seq* dengan memanfaatkan informasi gen penanda sebagai pengetahuan awal. Menggunakan 12 *dataset* scRNA-seq dengan label sel yang diketahui, rssNMF menunjukkan akurasi klaster yang lebih tinggi (rata-rata *Adjusted Rand Index* hingga 0,92) dibandingkan NMF standar, KMeans, dan *Hierarchical Clustering*. Model ini juga berhasil mengidentifikasi subkelas sel dan proses biologis laten, membuktikan bahwa integrasi pengetahuan biologis awal pada NMF dapat meningkatkan akurasi klaster dan interpretabilitas biologis.

Penelitian terdahulu mengenai *topic modeling*, menunjukkan bahwa LDA (*Latent Dirichlet Allocation*) menjadi metode dominan dalam *topic modeling*. Puspita dkk. (2024) misalnya, menerapkan LDA pada data pemberitaan online terkait merek Somethinc dan mendapatkan jumlah topik optimal sebanyak 6 dengan nilai *coherence* 0,404. Destarani dkk. (2019) juga menggunakan LDA untuk menganalisis keluhan warga Denpasar. Hasil dari model LDA menemukan topik tren pada topik ke-4 dengan nilai probabilitas 0,055 dan 78,95% topik dalam model tersebut dapat diinterpretasi dengan baik. Demikian pula, Tondang dkk. (2023) menerapkan LDA pada ulasan aplikasi BNI *Mobile Banking*, BCA Mobile, dan BRImo, berhasil mengelompokkan tema yang dominan dalam ulasan pengguna terhadap ketiga aplikasi tersebut, sehingga dapat memberikan pemahaman yang lebih mendalam mengenai preferensi pengguna, kualitas layanan masing-masing aplikasi, serta implikasi teknologi dalam konteks layanan perbankan *mobile*. Hasil penelitian menunjukkan adanya perbedaan karakteristik dalam topik-topik yang muncul pada setiap aplikasi, yang mencerminkan variasi dalam kualitas layanan yang ditawarkan.

Penelitian oleh Mustahidah (2021) menggunakan model PLSA (*Probabilistic Latent Semantic Analysis*) dan LDA dalam mengekstrak topik dari ulasan hotel. Penelitian ini mengklasifikasikan ulasan positif dan negatif untuk tiga hotel bintang empat dan mengevaluasi kualitas klasterisasi dengan *Calinski-Harabasz*. Hasilnya menunjukkan bahwa LDA secara konsisten memperoleh nilai indeks *Calinski-Harabasz* lebih tinggi dibandingkan PLSA pada setiap dataset, baik ulasan positif maupun negatif.

Selanjutnya, penelitian Mohammed & Al-Augby (2020) melakukan perbandingan langsung antara LDA dan LSA (*Latent Semantic Analysis*) menggunakan korpus teks ilmiah

yang terdiri dari 100 buku dan berisi sekitar 1 juta kata. Hasil penelitian ini memperlihatkan bahwa LDA mencapai nilai koherensi yang lebih tinggi dengan nilai 0,54846 pada topik ke-20. Artinya, topik-topik yang dihasilkan LDA lebih konsisten secara semantik dalam konteks tersebut. Ringkasnya, penelitian ini menyimpulkan LDA memberikan hasil pemodelan topik yang lebih baik daripada LSA untuk klasifikasi dokumen teks tidak terstruktur.

Penelitian oleh Nugumanova dkk. (2022) menerapkan lima algoritma Non-negative Matrix Factorization (NMF) yang berbeda pada *dataset* ACTER, yang berisi dokumen-dokumen khusus bidang tertentu yang telah dianotasi untuk keperluan terminologi. Penelitian ini dilatarbelakangi oleh keinginan untuk mengembangkan metode *unsupervised* dalam mengekstraksi istilah bidang tertentu tanpa memerlukan anotasi manual. Mereka melakukan eksperimen dengan memvariasikan jumlah topik (dari 5 hingga 50) dan jumlah kata teratas per topik (dari 5 hingga 30), serta membandingkan performa NMF dengan metode lain yang digunakan dalam *TermEval* 2020 *Shared Task*. Konfigurasi terbaik dari NMF berhasil mencapai nilai F1-*score* sebesar 0,47, berada tepat di bawah dua metode *supervised* yang lebih kompleks, dan menunjukkan bahwa NMF mampu mengekstraksi istilah relevan bidang tertentu secara andal dengan tingkat pengawasan yang minimal.

Penelitian lain yang dilakukan oleh Wu dkk. (2020) menerapkan model *robust semi-supervised NMF* (rssNMF) untuk mengelompokkan data RNA-seq sel tunggal (*single-cell RNA-seq*), dengan tujuan meningkatkan akurasi klasterisasi melalui pemanfaatan informasi gen penanda sebagai pengetahuan awal. Dengan menggunakan 12 *dataset* scRNA-seq yang telah memiliki label sel yang diketahui, model rssNMF menunjukkan akurasi klasterisasi yang lebih tinggi (rata-rata *Adjusted Rand Index* hingga 0,92) dibandingkan dengan NMF standar, KMeans, dan *Hierarchical Clustering*. Selain itu, model ini berhasil mengidentifikasi subkelas sel utama dan proses biologis laten, yang membuktikan bahwa integrasi pengetahuan biologis awal dengan NMF dapat meningkatkan performa klasterisasi sekaligus interpretasi biologisnya.

Tabel 2.2 Kelebihan dan Kekurangan Metode LDA, LSA, dan PLSA (Bagian 1)

Metode	Kelebihan	Kekurangan
LDA	Akurasi dan koherensi topik tinggi, mampu menangkap distribusi laten antar kata dan dokumen.	Sensitif terhadap parameter dan topik bisa tidak stabil antar eksekusi.
LSA	Efisien secara komputasi dan mampu menangkap hubungan semantik tidak langsung.	Tidak probabilistik, sehingga sulit menangani dokumen baru.

Tabel 2.2 Kelebihan dan Kekurangan Metode LDA, LSA, dan PLSA (Bagian 2)

Metode	Kelebihan	Kekurangan
PLSA	Representasi probabilistik kata-dokumen yang eksplisit, cocok untuk dataset kecil.	Rentan *overfitting* dan tidak bisa menggeneralisasi ke dokumen baru.
NMF	Efektif saat digunakan dengan fitur TF-IDF dan berguna untuk teks pendek dan jarang (*sparse*).	Sensitif terhadap representasi fitur dan tidak bersifat probabilistik secara bawaan.

Tabel 2.2 merupakan perbandingan kelebihan dan kekurangan dari beberapa metode topic modeling, yaitu LDA, LSA, PLSA, dan NMF. Metode *Latent Dirichlet Allocation* (LDA) merupakan pendekatan *probabilistic generative* yang paling banyak digunakan dalam pemodelan topik saat ini. Keunggulan utama dari LDA adalah kemampuannya dalam menghasilkan topik-topik yang koheren dan relevan pada berbagai domain data, baik berupa teks naratif umum maupun konten teknis. Menurut hasil studi oleh Puspita dkk. (2024), LDA mampu menghasilkan kelompok topik yang mendekati makna semantik sesungguhnya, terutama jika didukung oleh metrik koherensi seperti `C_V` dan `UMass`. Selain itu, LDA bersifat generatif, sehingga memungkinkan untuk memproyeksikan distribusi topik pada dokumen baru. Namun demikian, metode ini memiliki kekurangan dalam hal sensitivitas terhadap parameter. Nilai-nilai seperti jumlah topik (`num_topics`), serta parameter α dan β yang merepresentasikan distribusi *Dirichlet* untuk dokumen dan kata, sangat memengaruhi hasil akhir model. Ketika tidak dilakukan *tuning* dengan cermat, hasil yang diperoleh dapat berubah-ubah antar proses pelatihan (*non-deterministic output*) dan menurunkan stabilitas replikasi (Mustahidah, 2021).

Sementara itu, metode *Latent Semantic Analysis* (LSA) memiliki keunggulan dari sisi efisiensi komputasi. Pendekatan ini memanfaatkan teknik reduksi dimensi melalui *Singular Value Decomposition* (SVD), yang membuatnya cocok digunakan pada dataset berukuran besar dengan kebutuhan proses cepat. Selain itu, *LSA* dapat menangkap asosiasi semantik tidak langsung (*latent semantic structure*) antar istilah melalui matriks asosiasi kata-dokumen. Namun, kekurangan utama dari *LSA* adalah sifatnya yang non-probabilistik, sehingga tidak memiliki model eksplisit atas distribusi topik dalam dokumen. Hal ini menyebabkan kesulitan dalam memodelkan dokumen baru (*non-generative*), serta ketidakmampuan dalam mengestimasi kemungkinan kata baru muncul berdasarkan topik tertentu (Mohammed & Al-Augby, 2020).

Adapun *Probabilistic Latent Semantic Analysis* (PLSA) merupakan model awal yang memperkenalkan pendekatan probabilistik pada pemodelan topik. Keunggulan utama *PLSA*

terletak pada kemampuannya untuk memberikan distribusi topik yang eksplisit pada setiap dokumen, serta kecepatan estimasi parameter menggunakan algoritma *Expectation-Maximization* (EM) pada dataset berukuran kecil. Namun demikian, kelemahan utama *PLSA* adalah kerentanannya terhadap *overfitting*. Hal ini terjadi karena jumlah parameter model bertambah seiring banyaknya dokumen, tanpa adanya regularisasi (*Bayesian prior*) seperti yang dimiliki oleh *LDA*. Selain itu, model *PLSA* tidak bersifat generatif, sehingga tidak dapat digunakan untuk memproyeksikan distribusi topik pada dokumen baru (Mustahidah, 2021).

Metode *Non-negative Matrix Factorization* (NMF) dikenal mampu menghasilkan topik-topik yang berbeda dan dapat diinterpretasikan dengan baik, terutama saat menggunakan fitur TF-IDF. Dalam studi mengenai *tweet* berbahasa Urdu, NMF menunjukkan kinerja lebih baik dibandingkan metode lain saat menggunakan vektor fitur TF-IDF, sehingga cocok untuk menangani teks yang pendek dan jarang (Zoya dkk., 2021). Penelitian oleh (Nanyonga & Wild, 2025) mengidentifikasi tema-tema utama seperti kesalahan pilot, kegagalan mekanis, kondisi cuaca, dan kekurangan dalam pelatihan. Penggunaan NMF menghasilkan topik-topik yang jelas dan mudah dipahami, dengan skor koherensi sebesar 0,34, sedikit lebih rendah dari LDA, namun lebih tinggi dari PLSA.

Secara umum, berdasarkan hasil pengamatan dari lima publikasi ilmiah sebelumnya, metode LDA terbukti memberikan hasil yang lebih unggul dalam hal stabilitas dan koherensi topik dibandingkan dengan LSA, PLSA, maupun NMF. LDA Secara konsisten memberikan koherensi topik dan tingkat interpretasi tertinggi pada berbagai *dataset* teks berskala besar. Berbagai studi mengonfirmasi kinerja unggul LDA dalam mengekstraksi topik yang bermakna dan selaras dengan pemahaman manusia, menjadikannya pilihan yang paling tangguh dan andal untuk penerapan *topic modeling* tingkat lanjut.

Tabel 2.3 Bingkai Analisis Berdasarkan Pembuatan Ontologi (Bagian 1)

Sumber	Fokus Ontologi	Bahasa	Tools	Evaluasi	Deskripsi Singkat Penelitian
(Chatterjee dkk., 2021)	Manajemen gaya hidup sehat dan prediksi risiko kesehatan.	OWL	Protégé	SPARQL *query* dan HermiT *reasoner*	Studi ini memaparkan pengembangan sebuah ontologi berbasis OWL (UiA eHealth Ontology) untuk manajemen gaya hidup sehat, dengan fokus pada integrasi data kesehatan yang heterogen guna prediksi risiko dan pemberian rekomendasi yang dipersonalisasi.

Tabel 2.3 Bingkai Analisis Berdasarkan Pembuatan Ontologi (Bagian 2)

Sumber	Fokus Ontologi	Bahasa	Tools	Evaluasi	Deskripsi Singkat Penelitian
(Santosa dkk., 2023)	Pemodelan arsip Berita.	OWL	Protégé	SPARQL dan DL *query*	Pemodelan arsip berita menggunakan ontologi mampu merepresentasikan hubungan antar eleme seperti judul berita, tanggal terbit, kategori, kata kunci, ID admin, dan bagian halaman.
(Kurniawan dkk., 2024)	Pembangunan ontologi peraturan daerah tingkat dua di Kota Malang.	OWL	Protégé	Validasi oleh pakar	Ontologi membantu mitigasi risiko konstruksi dengan struktur formal berbasis hubungan risiko–penyebab–dampak.

Analisis terhadap penelitian terdahulu terkait ontologi yang disajikan pada Tabel 2.3, mayoritas studi menggunakan *Web Ontology Language* (OWL) sebagai bahasa utama karena kemampuannya dalam merepresentasikan pengetahuan secara eksplisit, formal, dan terstruktur. Bahasa OWL juga memungkinkan integrasi yang kuat dengan teknologi semantik dan mendukung reasoning otomatis untuk memverifikasi hubungan antar entitas.

Protégé menjadi perangkat lunak yang paling banyak digunakan dalam pengembangan ontologi. Tool ini menyediakan antarmuka yang mendukung visualisasi struktur konseptual serta kompatibel dengan plugin reasoning dan aturan tambahan (SWRL). Evaluasi ontologi dalam ketiga penelitian dilakukan melalui pendekatan berbasis SPARQL atau *reasoning* semantik. Studi oleh Chatterjee dkk. (2021) dan Santosa dkk. (2023) menggunakan SPARQL *query* dan HermiT *reasoner* untuk menguji validitas dan konsistensi hubungan antar entitas dalam ontologi. Sementara Kurniawan dkk. (2024) tetap menggunakan pendekatan berbasis OWL dan Protégé, namun evaluasi dilakukan melalui validasi pakar untuk memastikan kesesuaian ontologi dengan struktur pengetahuan yang diwakili. Konsistensi penggunaan OWL, Protégé, dan *SPARQL* menunjukkan bahwa kombinasi ini telah menjadi pendekatan yang kuat dan efektif dalam membangun serta mengevaluasi struktur ontologi yang semantik, valid, dan aplikatif di berbagai domain. Oleh karena itu, pendekatan serupa diadopsi dalam penelitian ini untuk menjamin keakuratan dan relevansi struktur ontologi yang dibangun.

Dari banyaknya rujukan peneliti terdahulu dalam keberhasilan penerapan metode *Latent Dirichlet Allocation* (LDA), hingga saat ini belum ditemukan studi yang secara khusus menerapkan LDA untuk *topic modeling* pada ulasan produk di platform Female Daily. Hal ini menunjukkan adanya celah penelitian, mengingat Female Daily merupakan salah satu forum

daring terbesar di Indonesia yang menjadi wadah utama opini konsumen terhadap produk kecantikan. Sebagian besar studi yang berfokus pada Female Daily masih terbatas pada analisis sentimen atau klasifikasi kepribadian, dan belum menjangkau eksplorasi topik laten dalam ulasan pengguna.

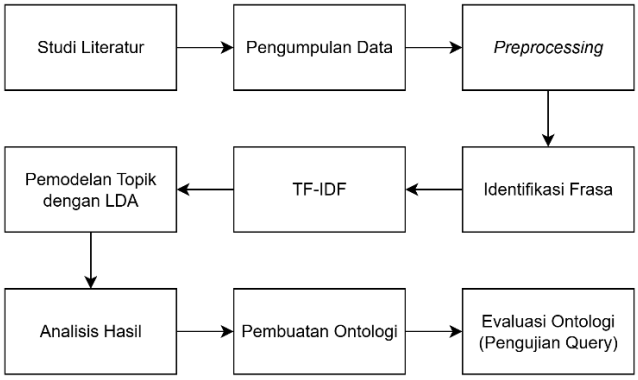
Maka dari itu penelitian ini akan menggunakan metode LDA dalam pembuatan *topic modeling* dengan perbedaan di bagian subjek penelitian yaitu komentar produk *skincare* yang meraih penghargaan dalam Female Daily *Best of Beauty Awards* pada tahun 2020 hingga 2024 pada *website* Female Daily yang akan dijelaskan di bagian metode penelitian dan pembahasan. Sehingga dapat dirumuskan masalah yang pertama, bagaimana menerapkan *topic modeling* menggunakan metode LDA pada judul tugas akhir mahasiswa. Kedua, bagaimana pembuatan ontologi terhadap model yang sudah dibuat dengan proses *topic modeling.*

Sebagai bentuk penguatan kontribusi kebaruan, penelitian ini tidak hanya menghasilkan pemodelan topik menggunakan metode *Latent Dirichlet Allocation* (LDA), tetapi juga melanjutkan tahapannya dengan pembuatan ontologi berbasis hasil pemodelan topik tersebut. Ontologi berfungsi untuk merepresentasikan struktur pengetahuan secara formal, sehingga konsep-konsep yang dihasilkan dari proses *topic modeling* dapat dikemas dalam bentuk skema yang terhubung secara semantik. Pembangunan ontologi divisualisasikan dalam format OWL menggunakan perangkat lunak Protégé yang memungkinkan entitas seperti produk *skincare*, *brand*, kandungan utama, jenis kulit, masalah kulit, dan manfaat produk diklasifikasikan serta dihubungkan dalam struktur konseptual yang dapat diakses secara sistematis. Berbeda dengan penelitian terdahulu yang umumnya berhenti pada tahap eksplorasi topik, penelitian ini memperluas cakupan pemanfaatan hasil LDA menjadi sumber pengetahuan eksplisit yang dapat dimanfaatkan untuk pengembangan sistem berbasis semantik atau analisis lanjut berbasis ontologi. Dengan demikian, integrasi antara LDA dan ontologi pada penelitian ini memberikan pendekatan terpadu yang belum banyak dikaji dalam konteks ulasan konsumen pada domain *skincare*.

BAB III
METODE PENELITIAN

3.1 Tahapan Penelitian

Tahapan penelitian diperlukan untuk memberikan gambaran yang jelas dan terstruktur mengenai proses penelitian. Diagram alir disusun untuk menggambarkan tujuh tahapan penelitian yang dilakukan secara sistematis yang ditunjukkan pada Gambar 3.1.



Gambar 3.1 Tahapan Penelitian

Tahapan pertama penelitian ini dimulai dengan melakukan studi literatur mengenai *topic modeling* dan pembuatan ontologi yang telah dilakukan oleh penelitian terdahulu lalu membandingkan tiap metode yang digunakan sebagai pertimbangan metode mana yang akan digunakan. Selanjutnya dilakukan proses pengumpulan data dari komentar produk *skincare* dalam format .xlsx yang digunakan sebagai data yang akan diolah. Setelah pengumpulan data, dilakukan *preprocessing* yang berisi *remove punctuation, case folding, tokenizing, stemming,* dan *stopword*. Setelah data bersih, tahap berikutnya yaitu transformasi data teks menjadi representasi numerik menggunakan TF-IDF. TF-IDF dilakukan untuk mengukur bobot atau pentingnya suatu kata (*term*) dalam suatu konten dari kumpulan dokumen. Nilai TF-IDF yang tinggi untuk suatu kata kunci dalam sebuah dokumen menunjukkan bahwa kata kunci tersebut memiliki peran penting dalam mewakili topik suatu dokumen. Setelah pembobotan kata, dilakukan pemodelan topik menggunakan LDA yang merupakan sebuah model probabilistik yang digunakan untuk mengidentifikasi topik-topik yang tersembunyi dalam sebuah korpus dokumen. Pemilihan LDA dibandingkan metode lain didasarkan pada keunggulan LDA dalam mengatasi permasalahan *overfitting* dan terbukti memberikan hasil yang lebih unggul dalam

memberikan koherensi topik dan tingkat interpretasi yang tinggi dibandingkan metode *topic modeling* lainnya.

Tahap selanjutnya yaitu pembuatan ontologi. Dalam penelitian ini, ontologi digunakan untuk merepresentasikan hubungan antar konsep yang dihasilkan dari analisis topik. Representasi ini bertujuan untuk mentransformasikan pengetahuan tematik hasil *topic modeling* dan data produk *skincare* menjadi bentuk formal berbasis OWL, sehingga dapat dimanfaatkan untuk kebutuhan analisis lanjutan di domain *skincare*. Setelah ontologi dibuat, akan dilakukan evaluasi menggunakan SPARQL *query* untuk merepresentasikan model ontologi produk *skincare* yang dibangun dengan pertanyaan-pertanyaan yang diajukan terkait dengan domain produk *skincare.* Alur penelitian ini bersifat linear mulai dari pengumpulan data hingga pembuatan ontologi. Tidak terdapat proses iterasi atau *feedback loop* formal setelah tahap evaluasi, sehingga setiap tahap dilakukan sekali sesuai urutan yang telah ditentukan.

3.2 Pengumpulan Data

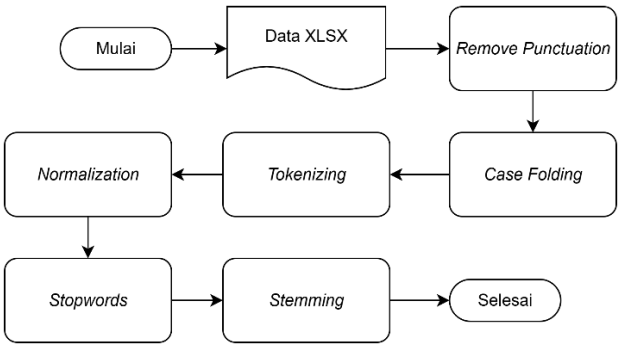
Proses pengumpulan data dilakukan dengan memanfaatkan ekstraksi data atau *web scraping* data dari salah satu forum *online* yaitu Female Daily dengan mengambil data komentar dari produk *skincare*. *Web scraping* merupakan teknik pengambilan data tertentu secara semi-terstruktur pada sebuah halaman web. Komentar yang akan dilakukan *scraping* adalah data dari komentar produk *skincare* yang mendapat penghargaan dalam *Best of Beauty Awards* di Female Daily dengan periode tahun 2020 hingga 2024 dan rentang waktu komentar produk antara 2015 hingga 2024. Jenis data yang diambil adalah *username profile, rate,* tanggal, dan komentar. Kemudian data akan disimpan dalam format *.xlsx*.

Meskipun daftar produk ditentukan berdasarkan pemenang *Best of Beauty Awards* dengan periode tahun 2020 hingga 2024, rentang waktu komentar yang dikumpulkan diperluas hingga mencakup tahun 2015 sampai 2024. Hal ini didasari oleh temuan bahwa beberapa produk pemenang memiliki jumlah komentar yang relatif sedikit pada periode BoBA-nya saja. Oleh karena itu, untuk menjaga proporsionalitas dan representativitas data antar produk, seluruh komentar yang tersedia hingga tahun 2024 tetap diikutsertakan. Selain itu, sebagian besar produk pemenang umumnya telah beredar di pasaran sebelum tahun mereka memperoleh penghargaan, sehingga komentar dari tahun-tahun sebelumnya tetap relevan dalam menangkap persepsi dan pengalaman pengguna sejak awal peluncuran produk. Strategi ini memungkinkan analisis yang lebih komprehensif terhadap dinamika *electronic word of mouth* (e-WOM), serta

memberikan gambaran yang lebih utuh mengenai konsistensi maupun perubahan persepsi publik terhadap produk *skincare* dari waktu ke waktu.

3.3 *Preprocessing* Data

Pada tahap ini, *preprocessing* data merupakan langkah fundamental dalam analisis teks yang bertujuan untuk membersihkan dan mempersiapkan data teks sebelum diolah. Urutan tahapan dalam *preprocessing* dapat dilihat pada Gambar 3.2.



Gambar 3.2 Alur Tahapan *Preprocessing*

Tahap *preprocessing* berperan penting dalam mencegah nilai kelebihan data, data yang hilang, dan data yang tidak konsisten sehingga dapat menghambat proses analisis dan memberikan hasil yang tidak akurat. Berikut tahapan *preprocessing* yang dilakukan:

a. *Remove Punctuation*

Proses pembersihan elemen non-alfabet seperti URL, tag HTML, tautan HTTP, emoji, angka, simbol, tanda baca, karakter baris baru (*line break*), karakter Unicode, serta karakter tunggal yang berdiri sendiri. Elemen-elemen tersebut dianggap tidak memberikan kontribusi semantik yang berarti dalam analisis topik. Sebagai contoh, pada data komentar “Sukaaa banget sama toner ini, ajaib paraaah!!!” menjadi “Sukaa banget sama toner ini ajaib paraah” yang hanya menghasilkan karakter alfabet.

b. *Case Folding*

Teknik ini digunakan untuk mengubah seluruh huruf dalam dokumen menjadi huruf kecil atau huruf besar. Karakter selain huruf, seperti tanda baca atau simbol, dihilangkan dan dianggap sebagai *delimeter*. Sehingga tidak ada permasalahan mengenai perbedaan kapitalisasi dalam korpus teks, seperti kata “Awal” dan “awal” dianggap identik.

c. *Tokenizing*

Tahap *tokenizing* untuk memecah teks menjadi unit-unit lebih kecil, disebut token. Token dapat berupa kata, angka, simbol, tanda baca, atau entitas penting lainnya. *Tokenizing* bertujuan untuk mempermudah dalam mengolah teks lebih lanjut, seperti penerapan analisis statistik atau pemodelan bahasa.

d. *Normalization*

Tahap *normalization* dilakukan untuk mengubah kata-kata yang tidak normal pada data komentar menjadi sesuai dengan standar. Pada tahap ini, penulis membuat kamus lokal yang disimpan dalam format xlsx. Kamus tersebut bertujuan untuk mengubah struktur kata yang tidak standar KBBI (Kamus Besar Bahasa Indonesia), seperti kata singkatan pada kata "bgt" menjadi "banget", kata gaul (*slang*) pada kata "jeriwi" menjadi "jerawat", kata dengan kesalahan penulisan (*typo*) pada kata "wahak" menjadi "wajah", dan kata dengan karakter berulang pada kata "kereeennn" menjadi "keren".

e. *Stopwords*

Kata-kata umum yang sering muncul dalam jumlah besar dan dianggap tidak memiliki makna signifikan seperti “dan”, “yang”, “di”, dan lain-lain. Penggunaan *stopword* yaitu dengan menghapus kata-kata yang memiliki informasi rendah dari sebuah frasa, sehingga hanya terfokus pada kata-kata penting yang membawa informasi lebih penting.

f. *Stemming*

Proses mereduksi kata-kata menjadi bentuk dasarnya sesuai dengan KBBI (Kamus Besar Bahasa Indonesia). Hal ini dilakukan dengan menghilangkan imbuhan-imbuhan (awalan, akhiran, dan sisipan) yang melekat pada kata, sehingga kata-kata yang memiliki makna serupa dapat dikelompokkan menjadi satu, seperti kata "pemakaian" menjadi "pakai". Tahap ini dilakukan dengan menggunakan *library* Python *Sastrawi.*

3.4 *Term Frequency-Inverse Document Frequency* (TF-IDF)

Setelah proses *preprocessing,* langkah selanjutnya dalam analisis teks adalah menghitung bobot kata menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*). Nilai TF-IDF yang tinggi untuk suatu kata kunci dalam sebuah dokumen menunjukkan bahwa kata kunci tersebut memiliki peran penting dalam mewakili topik suatu dokumen. Metrik ini mempertimbangkan dua faktor utama yaitu *term frequency* (TF) atau

frekuensi kemunculan kata dalam suatu dokumen dan *inverse document frequency* (IDF) atau frekuensi dokumen yang mengandung kata tersebut.

3.5 *Topic Modeling* Menggunakan *Latent Dirichlet Allocation* (LDA)

Pada tahap *topic modeling*, LDA digunakan untuk menemukan topik tersembunyi dalam kumpulan dokumen, yang bekerja dengan memodelkan setiap dokumen sebagai campuran dari beberapa topik dan setiap topik merupakan distribusi probabilitas terhadap kata-kata yang ada dalam korpus. Topik-topik yang dihasilkan dari pengolahan data selanjutnya perlu dievaluasi menggunakan uji koherensi topik**.** Uji ini bertujuan untuk mengukur tingkat keterkaitan antar kata-kata yang ditemukan dalam setiap topik.

3.6 Analisis Hasil

Tahap ini dilakukan analisis topik untuk mengamati topik-topik utama yang diperoleh apakah sudah sesuai dengan hasil yang diinginkan atau perlu disesuaikan lebih lanjut. Setiap topik dianalisis berdasarkan kata-kata dengan bobot tertinggi yang muncul, untuk kemudian diinterpretasikan makna tematiknya secara kualitatif. Analisis ini mencakup pengelompokan makna dari kata kunci utama, pengamatan terhadap keterkaitan antar topik, serta validasi terhadap koherensi dan relevansi topik dengan data asli. Validasi dilakukan dengan membandingkan kumpulan kata kunci dengan beberapa komentar pengguna yang relevan, guna memastikan bahwa topik yang dihasilkan benar-benar mencerminkan diskusi atau opini dalam data.

3.7 Pembuatan Ontologi

Pada tahap pembuatan ontologi, hasil analisis akan digunakan sebagai dasar untuk membangun ontologi produk. Bahasa yang digunakan dalam pembuatan ontologi yaitu *Ontology Web Language* (OWL), untuk mempresentasikan ontologi. Perangkat lunak yang digunakan untuk pembuatan ontologi yaitu Protégé.

3.8 Evaluasi Ontologi

Tahap evaluasi dilakukan dengan cara pengujian dengan *query* menggunakan SPARQL *query* yang terdapat di perangkat lunak Protégé. Pernyataan yang diinginkan diubah dalam bentuk *query,* yang nantinya akan menampilkan hasil yang ada dalam ontologi yang telah dibuat.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Pengumpulan Data

Data yang dikumpulkan yaitu data komentar produk *skincare* yang mendapat penghargaan dalam *Best of Beauty Awards* di Female Daily dengan periode tahun 2020 hingga 2024. Tahap pengumpulan data melibatkan *library* Python yaitu *requests* yang digunakan untuk mengirim permintaan HTTP ke API (*Application Programming Interface*) Female Daily dan menerima responnya dalam format JSON. Data JSON yang diterima kemudian diproses menggunakan *library* json untuk mengekstrak informasi yang dibutuhkan, seperti nama pengguna, *rate*, tanggal, dan teks ulasan. *Library datetime* digunakan untuk memformat tanggal ulasan ke format yang lebih mudah dibaca yaitu format DD-MM-YYYY. Data yang telah diekstrak kemudian disimpan dalam format .xlsx. Gambar 4.1 merupakan perintah untuk mengimpor *library* Python yang dibutuhkan. Proses *scraping* data dari halaman *website* ulasan Female Daily diimplementasikan seperti pada Gambar 4.2

```
import json
import requests
import openpyxl
from datetime import datetime
```

Gambar 4.1 *Import library* Python

```
(1)   def main():
(2)       url = "https://reviews.femaledaily.com/api"
(3)       payload = {}
(4)       result = []
(5)
(6)       for count in range(0, 3):       # Ambil data dari halaman 1 - 3
(7)           headers = {
(8)             'key': 'client03-hsjuebbbjasqqqivAAAbr-XkQQjannesQBlob',
(9)             'version': '1.5',
(10)            'device': '3',
(11)            'endpoint':f'https://api.femaledaily.com/app/v1/reviews?p
(12)  roduct_slug=radian-c-double-active-
(13)  brighteningessence1?cat=&cat_id=0&age_range=
(14)  &skin_type=&skin_tone=&skin_undertone=&hair_texture=&hair_type=&ord
(15)  er=newest&page={count+1}',
(16)            'sec-fetch-dest': 'empty',
(17)            'sec-fetch-mode': 'cors',
(18)            'sec-fetch-site': 'same-origin',
(19)            'if-none-match': 'W/"523f0-Hl5Fix7TIwRGgbKSpu0SxIjk6CI"',
(20)            'te': 'trailers}
```

Gambar 4.2 Inisialisasi Variabel dan *Loop* untuk Mengambil Data

Proses pengambilan data dari *Female Daily API* dilakukan melalui serangkaian tahapan yang sistematis. Pertama, dimulai dengan membuat fungsi utama bernama `main()` yang bertugas menjalankan seluruh proses pengambilan data. Pada Gambar 4.2 pada baris pertama, fungsi `main()` didefinisikan sebagai wadah utama proses *scraping*. *Endpoint API* yang digunakan adalah `https://reviews.femaledaily.com/api`, sebagaimana dituliskan pada baris kedua. Selain itu, pada baris ketujuh hingga sebelas, disusun konfigurasi *header* yang memuat parameter autentikasi seperti `key`, versi API (`version`), serta *device ID*. Di dalam *header* tersebut juga tercantum parameter `endpoint`, yang pada baris kesebelas hingga kedua belas dituliskan dalam format *f-string*, sehingga bagian `page={count+1}` dapat menyesuaikan halaman komentar yang diminta secara dinamis.

Pemanggilan data dilakukan secara bertahap berdasarkan nomor halaman (*page-wise*), bukan secara langsung berdasarkan jumlah total komentar. Pada Gambar 4.2 baris keenam, digunakan struktur perulangan `for count in range(0, 3)` yang menunjukkan bahwa proses *scraping* dilakukan untuk tiga halaman pertama komentar, yaitu halaman 1 hingga 3. Nilai `range(0, 3)` akan menghasilkan iterasi dengan nilai `count` dari 0 hingga 2, lalu pada baris kedua belas bagian `page={count+1}` menghasilkan nilai 1 hingga 3, yang merepresentasikan nomor halaman aktual. Pembatasan ini dilakukan karena produk yang diambil sebagai contoh hanya memiliki komentar hingga tiga halaman. Namun, dalam implementasi sebenarnya, kode ini dimodifikasi secara dinamis untuk menyesuaikan jumlah halaman yang tersedia per produk. Oleh karena itu, pendekatan ini bersifat fleksibel dan memungkinkan pengambilan data secara menyeluruh hingga halaman terakhir yang tersedia, bukan hanya terbatas pada komentar awal. Strategi ini digunakan agar hasil *scraping* lebih representatif terhadap persebaran opini pengguna di seluruh halaman komentar.

Setelah *request* berhasil dilakukan menggunakan pustaka `requests`, respons dari *API* yang berbentuk *JSON* dikonversi menjadi struktur *dictionary* Python menggunakan metode `.json()`. Meskipun bagian ini tidak terlihat pada potongan Gambar 4.2 karena berada setelah baris kedua puluh, proses ini penting untuk mengubah data dari format *JSON* ke dalam struktur Python agar dapat diakses dan diproses lebih lanjut. Data *review* yang diterima kemudian diproses satu per satu dalam *loop* terpisah. Setiap data *review* mengalami transformasi format, khususnya pada bagian *timestamp* yang dikonversi dari format ISO (`YYYY-MM-DDThh:mm:ss.µµµZ`) menjadi format yang lebih mudah dibaca (`DD-MM-YYYY HH:MM AM/PM`) menggunakan fungsi `datetime.strptime()` dan `strftime()`.

Setiap *review* yang diproses disusun dalam struktur `dictionary` yang terstandarisasi, mencakup empat *field* utama, yaitu *username profile* (diambil dari `data['user']['username']`), *rating produk* (`data['rating']`), tanggal *review* yang telah diformat, dan teks ulasan (`data['text']`). Hasil konversi ini kemudian dimasukkan ke dalam list `result`, sebagaimana dideklarasikan sebelumnya pada baris keempat dalam Gambar 4.2. Proses ini dilakukan secara berulang untuk setiap *review* dalam setiap halaman yang diakses, menghasilkan kumpulan data yang terstruktur dan siap untuk dianalisis lebih lanjut atau disimpan ke dalam format yang diinginkan.

```
(1)   if result:
(2)       workbook = openpyxl.Workbook()
(3)       worksheet = workbook.active
(4)       headers = list(result[0].keys())
(5)       worksheet.append(headers)
(6)
(7)       for row in result:
(8)           worksheet.append([row[key] for key in headers])
(9)       workbook.save('C:/Users/hp/Documents/SELENIUM/data_review_2024
(10)           /2024_TonerEssence_Laneige    Radian-C    Double    Active
(11)  BrighteningEssence.xlsx')
(12)           print('finish')
(13)      else:
(14)           print("Tidak ada data yang diperoleh")
(15)  if_name == '_main__':
(16)      main()
```

Gambar 4.3 Penyimpanan Data ke dalam Format .xlsx

Setelah proses pengambilan data selesai, dilakukan tahap penyimpanan data ke dalam format .xlsx dengan menggunakan *library* openpyxl. Proses yang disajikan pada Gambar 4.3 diawali dengan pemeriksaan kondisi terhadap *list* `result` untuk memastikan bahwa data yang akan disimpan tidak kosong, dengan menggunakan *conditional statement* `if result`. Pemeriksaan ini merupakan langkah validasi yang penting untuk menghindari pembuatan *file* Excel kosong atau *error* dalam proses penyimpanan. Dalam penelitian ini, proses *scraping* dilakukan secara terpisah untuk setiap produk *skincare*. Penamaan file dilakukan secara langsung pada baris kesembilan hingga kesebelas melalui perintah `workbook.save()`, dengan mencantumkan nama produk secara eksplisit dalam nama file `.xlsx`. Hal ini bertujuan untuk memudahkan proses identifikasi file berdasarkan nama produk yang di-*scrape*. Penelitian ini mendapatkan data komentar sebanyak 73 *file* .xlsx produk *skincare* dan terkumpul sebanyak 66.521 data komentar. Data lengkap hasil *scraping* dapat dilihat pada Lampiran A.

Tabel 4.1 Hasil *Scraping* Data

Username Profile	Rate	Date	Reviews
windyh	5	10/06/2023 15:33:00	Sunscreen gel yang nyaman di aku. Ringan, mudah menyerap, no whitecast, no peeling. Aman buat oily skin. Pertama kali pakai sudah cocok sama sunscreen ini. Ketagihan sampai sekarang sudah habis 3 tube. Nyaman banget dipakai setiap hari. Next bakal beli lagi.
sakinahapr	5	04/05/2024 06:17:00	Awal pemakaian muncul merah-merah, tapi alhamdulillah setelah pemakaian lebih dari satu tahun ini kulit atau skin barrier saya membaik, dan bekas-bekas atau noda-noda di wajah berangsur menghilang. Saya juga menggunakan berbagai produk skintific lainnya.
syifawiyana	5	03/05/2024 05:53	Azarine best banget, harganya murah tapi kualitasnyaa so pasti gak murahan. Awet banget kalian harus purchase dan terus re purchase produk azarine. Beli produk ini sekalian juga dukung produk lokal ha guys!! Go gurll glow with azarine luv

Contoh hasil dari pengambilan data dapat dilihat pada Tabel 4.1. Tabel ini memuat empat atribut utama yang berhasil diambil dari platform Female Daily, yaitu '*Username Profil*', '*Rate*', '*Date*', dan '*Reviews*'. Kolom '*Username Profil*' berisi nama pengguna yang memberikan ulasan terhadap produk tersebut. Kolom '*Rate*' menunjukkan skor penilaian (*rating*) yang diberikan oleh pengguna terhadap produk tersebut, biasanya dalam skala 1 hingga 5, di mana nilai 1 menunjukkan ketidakpuasan dan nilai 5 menunjukkan kepuasan tertinggi terhadap produk. Kolom '*Date*' merekam waktu dan tanggal saat ulasan tersebut dipublikasikan oleh pengguna. Kolom '*Reviews*' memuat isi komentar atau ulasan dari pengguna yang biasanya berisi pengalaman, pendapat, serta efek yang dirasakan setelah penggunaan produk. Teks *reviews* hasil *scraping* berisi data ulasan dan beberapa *noise* berupa elemen non-alfanumerik seperti angka, simbol khusus, dan tanda baca. Hasil ini membutuhkan tahap *preprocessing* agar mendapatkan data yang sempurna dan konsisten sebelum data dianalisis pada tahap selanjutnya.

4.2 Implementasi *Preprocessing*

Tahap *preprocessing* dilakukan untuk membersihkan dan mengorganisasikan data mentah agar siap untuk dianalisis lebih lanjut. Proses ini penting untuk menghilangkan ketidakkonsistenan, nilai yang hilang, atau data yang tidak relevan yang dapat menghambat akurasi hasil analisis.

4.2.1 Remove Punctuation

Tahap pertama yang dilakukan yaitu *remove punctuation* yang bertujuan untuk membersihkan data selain alfabet. Sehingga menghasilkan korpus teks yang lebih murni dan konsisten yang hanya terdiri dari karakter alfabet. Pada Gambar 4.4 menunjukkan implementasi dari *remove punctuation.*

```
import re
import string
import nltk

def remove_URL(text):
    url = re.compile(r'https?://\S+')
    return url.sub(r'', text)           # hapus URL

def remove_html(text):
    html = re.compile(r'<.*?>')
    return html.sub(r'', text)          # hapus html

def remove_http(text):
    http = re.compile(r'http\S+')
    return http.sub(r'', text)          # hapus http

def remove_emoji(text):
    return re.sub(r"[\U00010000-\U0010FFFF\u2600-\u27ff\u2300-\u23ff\u2500-\u257f\
     u2000-\u206f]+", '', text)         # hapus emoji

def remove_number(text):
    number = re.compile(r'\d+')
    return number.sub(r'', text)        # hapus number

def remove_punctuation(text):
    punctuation = re.compile(r'[^\w\s]')
    return punctuation.sub(r'', text)   # hapus punctuation

def remove_symbol(text):
    symbol = re.compile(r'[^\w\s]')
    return symbol.sub(r'', text)        # hapus symbol

def remove_linebreak(text):
    linebreak = re.compile(r'\n')
    return linebreak.sub(r'', text)     # hapus line break

def remove_single_char(text):
    single_char = re.compile(r'\b[a-aZ-Z]\b')
    return single_char.sub(r'', text)   # hapus single character

def remove_superscripts(text):
    pattern = r'\d+\^-\d+'               # hapus superscript
    text = re.sub(pattern, '', text)
    return text

def remove_repeated_chars(text):
    return re.sub(r'(.)\1{2,}', r'\1\1', text)  # Simpan dua kemunculan

def clean_unicode_symbols(text):           # hapus unicode
    text = re.sub(r'[²^°]', '', text)
    text = ''.join(
        c for c in unicodedata.normalize('NFD', text)
        if unicodedata.category(c) != 'Mn')
    return text
```

```
def join_words(text):
    words = text.split()
    return ' '.join(words)

df['clean_reviews'] = df['reviews'].fillna('') # Handle NaN values
df['clean_reviews'] = df['clean_reviews'].apply(lambda x: remove_html(x))
df['clean_reviews'] = df['clean_reviews'].apply(lambda x: remove_http(x))
df['clean_reviews'] = df['clean_reviews'].apply(lambda x: remove_emoji(x))
df['clean_reviews'] = df['clean_reviews'].apply(lambda x: remove_number(x))
df['clean_reviews'] = df['clean_reviews'].apply(lambda x: remove_punctuation(x))
df['clean_reviews'] = df['clean_reviews'].apply(lambda x: remove_symbol(x))
df['clean_reviews'] = df['clean_reviews'].apply(lambda x: remove_linebreak(x))
df['clean_reviews'] = df['clean_reviews'].apply(lambda x: remove_superscripts(x))
df['clean_reviews'] = df['clean_reviews'].apply(lambda x: remove_single_char(x))
df['clean_reviews'] = df['clean_reviews'].apply(lambda x: remove_repeated_chars(x))
df['clean_reviews'] = df['clean_reviews'].apply(clean_unicode_symbols)
df['clean_reviews'] = df['clean_reviews'].apply(lambda x: join_words(x))
```

Gambar 4.4 Kode Program *Remove Punctuation*

Implementasi kode pada Gambar 4.4 dilakukan dengan menyusun sejumlah fungsi pembersihan berbasis pustaka `re` (*regular expression*) dan `unicodedata` untuk memproses teks ulasan pengguna secara menyeluruh. Dalam proses ini, diterapkan serangkaian fungsi untuk menghapus berbagai karakter non-alfabet, seperti URL, *tag* HTML, tautan HTTP, emoji, angka, simbol, tanda baca, karakter baris baru (*line break*), karakter *unicode*, serta karakter tunggal yang berdiri sendiri, yang ditandai dengan *highlight* teks berwarna kuning. Selain itu, proses ini juga menghapus karakter yang terduplikasi lebih dari dua kali untuk menjaga konsistensi bentuk kata, serta melakukan normalisasi dengan menghilangkan *unicode superscript* yang tidak relevan. Seluruh fungsi di atas diterapkan secara bertahap pada kolom `reviews` yang sebelumnya telah ditangani nilai kosongnya (`fillna('')`). Hasil dari setiap fungsi disimpan kembali dalam kolom `clean_reviews` menggunakan metode `.apply()` secara berurutan. Hasil akhir dari tahapan ini dapat dilihat pada kolom *Clean_Text* dalam Tabel 4.2, yang menyajikan ulasan yang sudah dibersihkan, sehingga hanya menghasilkan teks yang lebih murni dan konsisten yang hanya terdiri dari karakter alfabet.

Tabel 4.2 Hasil Data Setelah Dilakukan *Remove Punctuation* (Bagian 1)

Reviews	***Clean_Text***
Sunscreen gel yang nyaman di aku. Ringan, mudah menyerap, no whitecast, no peeling Aman buat oily skin. Pertama kali pakai sudah cocok sama sunscreen ini. Ketagihan sampai sekarang sudah habis 3 tube. Nyaman banget dipakai setiap hari. Next bakal beli lagi.	Sunscreen gel yang nyaman di aku Ringan mudah menyerap no whitecast no peeling Aman buat oily skin Pertama kali pakai sudah cocok sama sunscreen ini Ketagihan sampai sekarang sudah habis tube Nyaman banget dipakai setiap hari Next bakal beli lagi

Tabel 4.2 Hasil Data Setelah Dilakukan *Remove Punctuation* (Bagian 2)

Reviews	***Clean_Text***
Awal pemakaian muncul merah-merah, tapi alhamdulillah setelah pemakaian lebih dari satu tahun ini kulit atau skin barrier saya membaik, dan bekas-bekas atau noda-noda di wajah berangsur menghilang. Saya juga menggunakan berbagai produk skintific lainnya.	Awal pemakaian muncul merah merah tapi alhamdulillah setelah pemakaian lebih dari satu tahun ini kulit atau skin barrier saya membaik dan bekas bekas atau noda noda di wajah berangsur menghilang Saya juga menggunakan berbagai produk skintific lainnya
Azarine best banget, harganya murah tapi kualitasnyaa so pasti gak murahan. Awet banget kalian harus purchase dan terus re purchase produk azarine. Beli produk ini sekalian juga dukung produk lokal ha guys!! Go gurll glow with azarine luv	Azarine best banget harganya murah tapi kualitasnyaa so pasti gak murahan Awet banget kalian harus purchase dan terus re purchase produk azarine Beli produk ini sekalian juga dukung produk lokal ha guys Go gurll glow with azarine luv

4.2.2 Case Folding

Setelah tahap pembersihan data, dilakukan penyamaan semua bentuk huruf dari keseluruhan teks menjadi huruf kecil. Tahap *case folding* mengimplementasikan *method* Python yaitu `lower()`. Hasil dari tahap ini dapat dilihat pada Tabel 4.3. Pada kolom *Case_Folding* menyajikan komentar yang sudah melalui tahap *case folding.*

Tabel 4.3 Hasil Data Setelah Dilakukan *Case Folding*

Clean_Text	***Case_Folding***
Sunscreen gel yang nyaman di aku Ringan mudah menyerap no whitecast no peeling Aman buat oily skin Pertama kali pakai sudah cocok sama sunscreen ini Ketagihan sampai sekarang sudah habis tube Nyaman banget dipakai setiap hari Next bakal beli lagi	sunscreen gel yang nyaman di aku ringan mudah menyerap no whitecast no peeling aman buat oily skin pertama kali pakai sudah cocok sama sunscreen ini ketagihan sampai sekarang sudah habis tube nyaman banget dipakai setiap hari next bakal beli lagi
Awal pemakaian muncul merah merah tapi alhamdulillah setelah pemakaian lebih dari satu tahun ini kulit atau skin barrier saya membaik dan bekas bekas atau noda noda di wajah berangsur menghilang Saya juga menggunakan berbagai produk skintific lainnya	awal pemakaian muncul merah merah tapi alhamdulillah setelah pemakaian lebih dari satu tahun ini kulit atau skin barrier saya membaik dan bekas bekas atau noda noda di wajah berangsur menghilang saya juga menggunakan berbagai produk skintific lainnya
Azarine best banget harganya murah tapi kualitasnyaa so pasti gak murahan Awet banget kalian harus purchase dan terus re purchase produk azarine Beli produk ini sekalian juga dukung produk lokal ha guys Go gurll glow with azarine luv	azarine best banget harganya murah tapi kualitasnyaa so pasti gak murahan awet banget kalian harus purchase dan terus re purchase produk azarine beli produk ini sekalian juga dukung produk lokal ha guys go gurll glow with azarine luv

4.2.3 Tokenizing

Tahap selanjutnya yaitu dilakukan *tokenizing* yang bertujuan untuk memisahkan kata pada sebuah teks menjadi satuan kata atau biasa disebut “token”. Implementasi *tokenizing*

menggunakan fungsi `word_tokenize` dari *library* Python yaitu `nltk` yang dapat dilihat pada Gambar 4.5. Pada baris keempat hingga kelima, didefinisikan fungsi `tokenize_nltk(text)` yang akan mengembalikan hasil pemrosesan `word_tokenize(text)`, yaitu daftar token dari teks masukan. Kemudian, pada baris keenam, fungsi `apply()` diterapkan pada kolom `'case_folding'` dari *dataframe* `df`, dengan menggunakan *lambda function* untuk menerapkan fungsi `nltk.tokenize.word_tokenize(x)` ke setiap elemen. Hal ini memungkinkan proses tokenisasi dilakukan secara otomatis untuk setiap baris teks. Proses ini menghasilkan daftar kata individual yang disimpan pada kolom baru `'tokenizing'`. Hasil dari proses ini dapat dilihat pada Tabel 4.4, yaitu pada kolom *Tokenizing*.

```
(1)   import nltk
(2)   from nltk.tokenize import word_tokenize
(3)
(4)   def tokenize_nltk(text):
(5)       return word_tokenize(text)
(6)   df['tokenizing']=df['case_folding'].apply(lambdax:nltk.tokenize.wor
(7)   d_tokenize(x))
```

Gambar 4.5 Kode Program *Tokenizing*

Tabel 4.4 Hasil Data Setelah Dilakukan *Tokenizing*

Case_Folding	***Tokenizing***
sunscreen gel yang nyaman di aku ringan mudah menyerap no whitecast no peeling aman buat oily skin pertama kali pakai sudah cocok sama sunscreen ini ketagihan sampai sekarang sudah habis tube nyaman banget dipakai setiap hari next bakal beli lagi	['sunscreen', 'gel', 'yang', 'nyaman', 'di', 'aku', 'ringan', 'mudah', 'menyerap', 'no', 'whitecast', 'no', 'peeling', 'aman', 'buat', 'oily', 'skin', 'pertama', 'kali', 'pakai', 'sudah', 'cocok', 'sama', 'sunscreen', 'ini', 'ketagihan', 'sampai', 'sekarang', 'sudah', 'habis', 'tube', 'nyaman', 'banget', 'dipakai', 'setiap', 'hari', 'next', 'bakal', 'beli', 'lagi']
awal pemakaian muncul merah merah tapi alhamdulillah setelah pemakaian lebih dari satu tahun ini kulit atau skin barrier saya membaik dan bekas bekas atau noda noda di wajah berangsur menghilang saya juga menggunakan berbagai produk skintific lainnya	['awal', 'pemakaian', 'muncul', 'merah', 'merah', 'tapi', 'alhamdulillah', 'setelah', 'pemakaian', 'lebih', 'dari', 'satu', 'tahun', 'ini', 'kulit', 'atau', 'skin', 'barrier', 'saya', 'membaik', 'dan', 'bekas', 'bekas', 'atau', 'noda', 'noda', 'di', 'wajah', 'berangsur', 'menghilang', 'saya', 'juga', 'menggunakan', 'berbagai', 'produk', 'skintific', 'lainnya']
azarine best banget harganya murah tapi kualitasnyaa so pasti gak murahan awet banget kalian harus purchase dan terus re purchase produk azarine beli produk ini sekalian juga dukung produk lokal ha guys go gurll glow with azarine luv	['azarine', 'best', 'banget', 'harganya', 'murah', 'tapi', 'kualitasnyaa', 'so', 'pasti', 'gak', 'murahan', 'awet', 'banget', 'kalian', 'harus', 'purchase', 'dan', 'terus', 're', 'purchase', 'produk', 'azarine', 'beli', 'produk', 'ini', 'sekalian', 'juga', 'dukung', 'produk', 'lokal', 'ha', 'guys', 'go', 'gurll', 'glow', 'with', 'azarine', 'luv']

4.2.4 Normalization

Proses *normalization* bertujuan untuk mentransformasi kata-kata yang tidak sesuai dengan standar ke dalam bentuk yang baku. Pada Gambar 4.6, fungsi `get_lookup()` digunakan untuk memuat kamus normalisasi dari file *Excel* bernama kamus.xlsx. Pada baris ketiga hingga kelima, file dibaca menggunakan fungsi `pd.read_excel()` dan diubah menjadi format *dictionary* menggunakan metode `to_dict('records')`. Selanjutnya, pada baris keenam hingga kedelapan, *dictionary* tersebut dibangun dengan memetakan setiap entri kata tidak baku (`'tidak_baku'`) ke kata bakunya (`'kata_baku'`), sehingga proses substitusi dapat dilakukan secara otomatis.

Pada baris kesepuluh, fungsi `get_lookup()` dipanggil dan hasilnya disimpan ke dalam variabel `lookup`, yang akan digunakan sebagai referensi dalam proses normalisasi. Pada baris kedua belas, proses iterasi dilakukan terhadap seluruh token, dan setiap token akan diganti dengan bentuk bakunya berdasarkan isi *dictionary* `lookup` menggunakan metode `lookup.get(token, token)`. Metode ini memastikan bahwa jika suatu token tidak ditemukan dalam kamus, maka token tersebut akan tetap dipertahankan dalam bentuk aslinya.

Setelah proses substitusi selesai, pada baris ketiga belas, token hasil normalisasi dihitung frekuensinya menggunakan `FreqDist()` dari pustaka *nltk*, sehingga hasil akhir berupa token-token unik yang telah dinormalisasi. Terakhir, pada baris ketujuh belas, fungsi `apply()` digunakan untuk menerapkan proses normalisasi ini pada setiap entri dalam kolom `'tokenizing'` dan menyimpannya ke dalam kolom baru bernama `'normalization'`.

```
(1)  from nltk.probability import FreqDist
(2)
(3)  def get_lookup():
(4)    df_lookup = pd.read_excel('C:/Users/hp/Documents/SELENIUM/kamus
(5)                .xlsx')
(6)    df_record = df_lookup.to_dict('records')
(7)      return {entry['tidak_baku']: entry['kata_baku'] for entry in
(8)      df_record}
(9)
(10) lookup = get_lookup()
(11)
(12) def formalize(tokens):
(13)     formalized = [lookup.get(token, token) for token in tokens]
(14)     freq_tokens = FreqDist(formalized)
(15)     return " ".join(freq_tokens.keys()
(16)
(17) df['normalization'] = df['tokenizing'].apply(formalize)
```

Gambar 4.6 Kode Program *Normalization*

Data komentar mengandung variasi bentuk kata tidak standar yang beragam. Beberapa bentuk kata tersebut memerlukan normalisasi seperti kata singkatan, *slang*, kesalahan pengetikan (*typo*), dan pengulangan karakter. Contoh perbaikan untuk setiap bentuk kata dapat dilihat pada Tabel 4.5, Tabel 4.6, Tabel 4.7, dan Tabel 4.8. Rincian lebih lanjut mengenai daftar kata non-standar dan hasil normalisasinya dilampirkan pada Lampiran B. Hasil implementasi dari tahap *normalization* diberi huruf *bold* dan dapat dilihat pada Tabel 4.9 yaitu pada kolom *Normalization.*

Tabel 4.5 Contoh Singkatan dan Perbaikannya

Singkatan	**Perbaikan**
udh	sudah
bgt	banget
pst	pasti
brp	berapa

Tabel 4.6 Contoh Kata *Slang* dan Perbaikannya

Slang Word	**Perbaikan**
canci	cantik
murce	murah
enakeun	enak
jeriwi	jerawat

Tabel 4.7 Contoh Kata Yang Tidak Sesuai Standar (*Typo*) dan Perbaikannya

Typo	**Perbaikan**
doank	doang
kurg	kurang
lembap	lembab
wajak	wajah

Tabel 4.8 Contoh Kata Dengan Karakter Berulang dan Perbaikannya

Singkatan	**Perbaikan**
kereeennn	keren
segerr	seger
baguus	bagus
enakkkkk	enak

Tabel 4.9 Hasil Data Setelah Dilakukan *Normalization*

Tokenizing	***Normalization***
['sunscreen', 'gel', 'yang', 'nyaman', 'di', 'aku', 'ringan', 'mudah', 'menyerap', 'no', 'whitecast', 'no', 'peeling', 'aman', 'buat', 'oily', 'skin', 'pertama', 'kali', 'pakai', 'sudah', 'cocok', 'sama', 'sunscreen', 'ini', 'ketagihan', 'sampai', 'sekarang', 'sudah', 'habis', 'tube', 'nyaman', 'banget', 'dipakai', 'setiap', 'hari', 'next', 'bakal', 'beli', 'lagi']	sunscreen gel yang nyaman di aku ringan mudah menyerap no whitecast no peeling aman buat oily skin pertama kali pakai sudah cocok sama sunscreen ini ketagihan sampai sekarang sudah habis tube nyaman banget pakai setiap hari next bakal beli lagi
['awal', 'pemakaian', 'muncul', 'merah', 'merah', 'tapi', 'alhamdulillah', 'setelah', 'pemakaian', 'lebih', 'dari', 'satu', 'tahun', 'ini', 'kulit', 'atau', 'skin', 'barrier', 'saya', 'membaik', 'dan', 'bekas', 'bekas', 'atau', 'noda', 'noda', 'di', 'wajah', 'berangsur', 'menghilang', 'saya', 'juga', 'menggunakan', 'berbagai', 'produk', 'skintific', 'lainnya']	awal pemakaian muncul merah merah tapi alhamdulillah setelah pemakaian lebih dari satu tahun ini kulit atau skin barrier saya membaik dan bekas bekas atau noda noda di wajah berangsur menghilang saya juga menggunakan berbagai produk skintific lainnya
['azarine', 'best', 'banget', 'harganya', 'murah', 'tapi', 'kualitasnyaa', 'so', 'pasti', 'gak', 'murahan', 'awet', 'banget', 'kalian', 'harus', 'purchase', 'dan', 'terus', 're', 'purchase', 'produk', 'azarine', 'beli', 'produk', 'ini', 'sekalian', 'juga', 'dukung', 'produk', 'lokal', 'ha', 'guys', 'go', 'gurll', 'glow', 'with', 'azarine', 'luv']	azarine best banget harga murah tapi kualitas so pasti **enggak murah** awet banget kalian harus purchase dan terus re purchase produk azarine beli produk ini sekalian juga dukung produk lokal ya guys go **girl** glow with azarine **love**

4.2.5 Stopwords

Setelah data komentar dinormalisasi ke dalam bentuk baku, tahapan *preprocessing* selanjutnya adalah penghapusan *stopwords*, yaitu kata-kata umum dengan frekuensi tinggi seperti “yang”, “dan”, atau “sudah” yang dinilai tidak memberikan kontribusi informatif dalam analisis teks. Tujuan utama dari proses ini adalah untuk mereduksi dimensi data dan meningkatkan efisiensi komputasi dengan memfokuskan analisis pada kata-kata yang lebih bermakna secara tematik. Untuk lebih meningkatkan kinerja model pemodelan topik,

penelitian melakukan proses kurasi manual terhadap kata-kata yang muncul dalam konteks spesifik ulasan *skincare*.

Pada tahap awal pengujian, daftar *stopwords* masih mengandalkan kombinasi dari pustaka standar bahasa Indonesia yang tersedia di laman https://www.kaggle.com/datasets/oswinrh/indonesian-stoplist dan Inggris yang tersedia di laman https://www.kaggle.com/datasets/rowhitswami/stopwords, ditambah dengan sejumlah kata tambahan yang dipilih berdasarkan frekuensi tinggi di dalam korpus. Namun, pada fase ini, seleksi kata dilakukan dengan fokus utama pada eliminasi kata-kata umum tanpa mempertimbangkan secara menyeluruh konteks semantik dalam domain *skincare*. Beberapa kata yang sebenarnya memiliki nilai informatif penting, seperti "tidak", "tapi", "mudah", dan sebagainya justru ikut dihapus karena dianggap terlalu sering muncul. Hal ini disebabkan karena penambahan awal pada daftar *stopwords* lebih diarahkan oleh sudut pandang teknis frekuensi kata, bukan dari sudut pandang konsumen atau pembaca domain. Dampak dari penghapusan ini terlihat pada hasil *stopwords* yang kehilangan kata-kata kunci utama dan menyebabkan topik yang dihasilkan menjadi kurang bermakna atau bersifat terlalu umum. Contoh representatif dari hasil ini ditunjukkan pada Tabel 4.10.

Tabel 4.10 Hasil Data Setelah Dilakukan *Stopwords* Awal

Normalization	***Stopwords***
sunscreen gel yang nyaman di aku ringan mudah menyerap no whitecast no peeling aman buat oily skin pertama kali pakai sudah cocok sama sunscreen ini ketagihan sampai sekarang sudah habis tube nyaman banget pakai setiap hari next bakal beli lagi	sunscreen gel ringan menyerap whitecast peeling oily
awal pemakaian muncul merah merah tapi alhamdulillah setelah pemakaian lebih dari satu tahun ini kulit atau skin barrier saya membaik dan bekas bekas atau noda noda di wajah berangsur menghilang saya juga menggunakan berbagai produk skintific lainnya	kulit skin barrier bekas berangsur menghilang skintific
azarine best banget harga murah tapi kualitas so pasti enggak murah awet banget kalian harus purchase dan terus re purchase produk azarine beli produk ini sekalian juga dukung produk lokal ya guys go girl glow with azarine love	azarine best murah kualitas murah awet lokal girl glow

Berdasarkan hasil tersebut, dilakukan peninjauan ulang terhadap daftar *stopwords*. Kata-kata penting yang berkontribusi dalam pembentukan topik, terutama yang sering digunakan dalam ulasan *skincare* dikeluarkan dari daftar *stopwords*, sementara kata-kata tambahan yang

bersifat umum dan tidak kontekstual ditambahkan secara manual. Melalui 10 iterasi evaluasi, dilakukan penyempurnaan daftar hingga menghasilkan daftar final berjumlah 28.796 stopwords.

Setiap iterasi dalam proses penyusunan *stopwords* melibatkan evaluasi terhadap hasil tokenisasi dan keluaran model topik sementara. Pada tahap ini, peneliti meninjau apakah masih terdapat kata-kata yang bersifat umum, tidak relevan, atau berfrekuensi tinggi namun tidak menyumbang makna tematik yang signifikan. Selain itu, dilakukan pengamatan terhadap topik-topik awal yang dihasilkan oleh model untuk mengidentifikasi apakah terdapat kata-kata yang secara semantik lemah atau menyimpang dari konteks domain. Temuan dari evaluasi ini digunakan untuk menyempurnakan daftar *stopwords* secara bertahap hingga mencapai hasil yang optimal. Hasil dari penggunaan daftar final ini menunjukkan peningkatan signifikan dalam kualitas token yang dipertahankan. Kata-kata penting tidak lagi terhapus, dan hasil topik modeling menjadi lebih tematik serta representatif terhadap pembahasan pengguna. Tabel 4.11 menampilkan perbandingan hasil tokenisasi menggunakan daftar *stopwords* final.

Tabel 4.11 Hasil Data Setelah Dilakukan *Stopwords* Final

Normalization	***Stopwords* Awal**	***Stopwords* Final**
awal pemakaian muncul merah merah tapi alhamdulillah setelah pemakaian lebih dari satu tahun ini kulit atau skin barrier saya membaik dan bekas bekas atau noda noda di wajah berangsur menghilang saya juga menggunakan berbagai produk skintific lainnya	kulit barrier bekas jerawat angsur hilang skintific	awal pemakaian muncul merah merah tapi setelah pemakaian lebih satu tahun kulit atau skin barrier saya membaik bekas bekas atau noda noda wajah berangsur menghilang saya juga menggunakan berbagai produk skintific lainnya

Perbandingan ini menunjukkan bahwa penggunaan *stopwords* yang terlalu agresif dapat menurunkan kualitas hasil analisis teks. Oleh karena itu, penyesuaian daftar *stopwords* secara iteratif berbasis pemahaman domain terbukti krusial dalam meningkatkan relevansi semantik model dan memperkuat validitas hasil pemodelan topik. Implementasi *stopwords* ditunjukkan pada Gambar 4.7. Pada baris pertama hingga ketiga, kode program memuat daftar *stopwords* dari file eksternal dengan menyimpan jalur file pada variabel `stop_words_file`, kemudian membaca isinya menggunakan perintah `file.read().splitlines()`. Hasil pembacaan tersebut disimpan ke dalam variabel `stop_words` sebagai daftar kata-kata yang akan dihapus dari teks. Selanjutnya, pada baris kelima hingga ketujuh, didefinisikan fungsi `remove_stopwords(text)` yang bertujuan untuk menghapus kata-kata dalam teks yang

termasuk ke dalam daftar *stopwords*. Sebelum melakukan proses pemrosesan, fungsi ini terlebih dahulu memeriksa apakah *input* `text` merupakan tipe data *string* menggunakan `isinstance(text, str)`. Jika bukan, maka fungsi akan langsung mengembalikan nilai teks tersebut tanpa perubahan.

Pada baris kedelapan hingga kesebelas, teks diproses dengan membagi kalimat menjadi daftar kata menggunakan fungsi `split()`, kemudian dilakukan penyaringan menggunakan *list comprehension* agar hanya kata-kata yang tidak termasuk dalam *stopwords* yang dipertahankan. Hasilnya disatukan kembali menjadi string menggunakan `' '.join()`. Pada baris ketigabelas, fungsi `remove_stopwords()` diterapkan pada kolom `"normalization"` dari *dataframe* `df`, dan hasilnya disimpan ke kolom baru bernama `"stopwords"` menggunakan metode `apply()`. Hasil dari proses ini ditampilkan pada Tabel 4.11, sedangkan daftar lengkap *stopwords* yang digunakan dapat dilihat pada Lampiran C.

```
(1)   stop_words_file = 'C:/Users/hp/Documents/SELENIUM/stopwords.txt'
(2)   with open(stop_words_file, 'r', encoding='utf-8') as file:
(3)        stop_words = file.read().splitlines()
(4)
(5)   def remove_stopwords(text):
(6)       if not isinstance(text, str):
(7)           return text
(8)       words = text.split()
(9)       filtered_words = [word for word in words if word.lower() not in
(10)                        stop_words]
(11)      return ' '.join(filtered_words)
(12)
(13) df["stopwords"] = df["normalization"].apply(remove_stopwords)
```

Gambar 4.7 Kode Program *Stopwords*

4.2.6 Stemming

Setelah penghapusan *stopwords*, dilakukan *stemming* untuk mengubah kata yang berimbuhan, awalan, sapaan, akhiran, atau kombinasi. Sehingga setiap kata dalam dokumen hanya berupa kata dasarnya sesuai Kamus Besar Bahasa Indonesia (KBBI). Implementasi *stemming* pada data komentar menggunakan pustaka *Sastrawi*, sebagaimana diperlihatkan pada Gambar 4.8. Pada baris pertama, dilakukan *import* terhadap kelas `StemmerFactory` dari modul `Sastrawi.Stemmer.StemmerFactory`. Selanjutnya, pada baris ketiga hingga keempat, objek `factory` dibentuk dari kelas `StemmerFactory()`, dan digunakan untuk menghasilkan objek `stemmer` melalui metode `create_stemmer()`.

Pada baris keenam hingga kedua belas, didefinisikan fungsi `stem_text(text)` yang bertugas melakukan *stemming* terhadap setiap kata dalam teks. Baris ketujuh berfungsi untuk memastikan bahwa masukan `text` merupakan tipe *string*; jika tidak, maka teks dikembalikan tanpa perubahan. Pada baris kesembilan, kalimat dipecah menjadi daftar kata menggunakan metode `split()`. Kemudian pada baris kesepuluh, setiap kata dalam daftar tersebut diproses menggunakan metode `stem()` dari objek `stemmer`. Hasil *stemming* untuk setiap kata disimpan dalam daftar `stemmed_words`. Pada baris kedua belas, daftar kata yang telah distem diubah kembali menjadi satu string menggunakan `' '.join()`. Hasil implementasi dari kode tersebut diterapkan ke dalam kolom `"stopwords"` pada *dataframe* dan disimpan sebagai kolom baru bernama `"stemming"`. Hasil implementasi dari tahap *stemming* diberi huruf *bold* dengan hasil yang ditampilkan pada Tabel 4.12.

```
(1)   from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
(2)
(3)   factory = StemmerFactory()
(4)   stemmer = factory.create_stemmer()
(5)
(6)   def stem_text(text):
(7)       if not isinstance(text, str):
(8)           return text
(9)       words = text.split()
(10)      stemmed_words = [stemmer.stem(word) for word in words]
(11)
(12)     return ' '.join(stemmed_words)
```

Gambar 4.8 Kode Program *Stemming*

Tabel 4.12 Hasil Data Setelah Dilakukan *Stopwords*

Stopwords	***Stemming***
sunscreen gel nyaman aku ringan mudah menyerap no whitecast no peeling aman buat oily skin pertama pakai sudah cocok sunscreen sampai sekarang sudah habis tube nyaman banget pakai setiap hari beli lagi	sunscreen gel nyaman aku ringan mudah **serap** no whitecast no peeling aman buat oily skin pertama pakai sudah cocok sunscreen sampai sekarang sudah habis tube nyaman banget pakai **tiap** hari beli lagi
awal pemakaian muncul merah merah tapi setelah pemakaian lebih satu tahun kulit atau skin barrier saya membaik bekas bekas atau noda noda wajah berangsur menghilang saya juga menggunakan berbagai produk skintific lainnya	awal **pakai** muncul merah merah tapi **telah** pakai lebih satu tahun kulit atau skin barrier saya **baik** bekas bekas atau noda noda wajah **angsur hilang** saya juga **guna bagai** produk skintific **lain**
azarine best banget harga murah tapi kualitas pasti enggak murah awet banget kalian harus purchase terus re purchase produk azarine beli produk juga dukung produk lokal glow azarine	azarine best banget harga murah tapi kualitas pasti enggak murah awet banget kalian harus purchase terus re purchase produk azarine beli produk juga dukung produk lokal glow azarine

4.3 Identifikasi Frasa

Setelah keseluruhan data melalui tahap *preprocessing*, selanjutnya dilakukan pembuatan model bigram dan trigram. Bigram didefinisikan sebagai pasangan dua kata yang memiliki frekuensi kemunculan tinggi dalam suatu dokumen, sedangkan trigram adalah rangkaian tiga kata dengan karakteristik serupa. Tahap identifikasi frasa ini bertujuan untuk mengekstraksi frasa kandidat berdasarkan pola kemunculan bersama (*co-occurency*) kata-kata dalam bentuk bigram dan trigram. Tahap ini menggunakan modul `Phrases` dan `Phraser` dari *library* Gensim. Implementasi algoritma pada tahap ini terlihat pada Gambar 4.9.

```
(1)   from gensim.models.phrases import Phrases, Phraser
(2)
(3)   # Buat model Bigram
(4)   bigram = Phrases(sentences,
(5)                    min_count=20,
(6)                    threshold=25,
(7)                    delimiter='_')
(8)   bigram_model = Phraser(bigram)
(9)
(10)  # Buat model Trigram
(11)  trigram = Phrases(bigram_model[sentences],
(12)                    min_count=20,
(13)                    threshold=25,
(14)                    delimiter='_')
(15)  trigram_model = Phraser(trigram)
(16)
(17)  def apply_phrases(text):
(18)      tokens = text.split()
(19)      bigram_tokens = bigram_model[tokens]
(20)      trigram_tokens = trigram_model[bigram_tokens]
(21)      return " ".join(trigram_tokens)
```

Gambar 4.9 Kode Program Algoritma Bigram dan Trigram

Proses yang dilakukan mencakup pencarian dan penggabungan kata-kata yang berdekatan menjadi unit frasa, sesuai dengan pola bigram dan trigram. Modul Phrases melakukan identifikasi pasangan kata (bigram) yang sering muncul bersamaan. Pada Gambar 4.9 pada baris kelima hingga enam, parameter `min_count=20` menentukan bahwa pasangan kata harus muncul setidaknya 20 kali agar dipertimbangkan, sementara `threshold=25` menentukan batas skor minimum yang harus dicapai oleh pasangan kata agar dianggap sebagai suatu frasa. Penentuan nilai ini didasarkan pada karakteristik korpus komentar yang dianalisis, yang terdiri dari 66.521 entri. Dengan jumlah data yang relatif besar, frekuensi kemunculan kata dan kombinasi kata cenderung lebih tinggi. Dengan parameter ini, frasa-frasa yang terbentuk cenderung memiliki keterkaitan semantik yang kuat dan kontekstual terhadap topik

pembahasan. Penggabungan frasa menggunakan tanda hubung *underscore*. Implementasinya dapat dilihat pada frasa "*skin_barrier*" dan "*re_purchase*" pada Tabel 4.13 berhasil teridentifikasi karena sering muncul dalam komentar dengan makna yang konsisten. Strategi ini diterapkan agar frasa-frasa yang dihasilkan benar-benar merepresentasikan istilah-istilah penting dalam diskusi seputar *skincare* dan dapat memperkuat hasil pemodelan topik di tahap selanjutnya. Setiap bigram dan trigram yang terbentuk dianggap sebagai frasa tunggal. Selanjutnya, frasa-frasa ini diintegrasikan ke dalam konteks kalimat utuh. Hasil dari tahapan ini diberi huruf *bold* dan dapat dilihat pada Tabel 4.13.

Tabel 4.13 Hasil Data Setelah Dilakukan Identifikasi Frasa

Data Komentar Setelah Tahap *Preprocessing*	**Hasil bigram dan trigram**
awal pakai muncul merah merah tapi telah pakai lebih satu tahun kulit atau skin barrier saya baik bekas bekas atau noda noda wajah angsur hilang saya juga guna bagai produk skintific lain	awal pakai muncul merah merah tapi telah pakai lebih satu tahun kulit atau **skin_barrier** saya baik bekas bekas atau noda noda wajah angsur hilang saya juga guna bagai produk skintific lain
azarine best banget harga murah tapi kualitas pasti enggak murah awet banget kalian harus purchase terus re purchase produk azarine beli produk juga dukung produk lokal glow azarine	azarine best banget harga murah tapi kualitas pasti enggak murah awet banget kalian harus purchase terus **re_purchase** produk azarine beli produk juga dukung produk lokal glow azarine

4.4 Pembobotan dengan TF-IDF

Tahap pembobotan dengan metode TF-IDF memanfaatkan *library* Gensim yang diawali dengan pembentukan kamus (*dictionary*) dari data yang telah diproses sebelumnya untuk merepresentasikan isi dari dari dokumen. Pembentukan kamus menggunakan modul *corpora* dengan fungsi `Dictionary()` yang dapat dilihat pada Gambar 4.10 baris kelima. Selanjutnya, dilakukan penyaringan kata-kata yang terlalu sering atau terlalu jarang muncul menggunakan metode `filter_extremes()` dari objek *dictionary*. Metode ini ditunjukkan pada Gambar 4.10 baris keenam dengan parameter `no_below=20` untuk mengeliminasi kata-kata yang muncul kurang dari 20 dokumen, sedangkan `no_above=0.5` untuk mengeliminasi kata-kata yang muncul di lebih dari 50% dokumen. Langkah ini bertujuan untuk mengurangi *noise* dan fokus pada kata-kata yang lebih informatif.

```
(1)   from gensim import corpora, models
(2)
(3)   def create_tf_idf_model(data):
(4)      # Buat dictionary representasi dari dokumen
(5)       dictionary = corpora.Dictionary(data)
(6)       dictionary.filter_extremes(no_below=20, no_above=0.5)
```

```
(7)
(8)      # Buat representasi bag of words
(9)       corpus = [dictionary.doc2bow(text) for text in data]
(10)
(11)     # Perhitungan bobot tiap kata
(12)     tfidf = models.TfidfModel(corpus)
(13)     corpus_tfidf = tfidf[corpus]
(14)
(15)    return dictionary, corpus_tfidf
(16)   dictionary, corpus tfidf = create tf idf model(data)
```

Gambar 4.10 Kode Program Pembobotan TF-IDF

Setelah membentuk *dictionary,* korpus diubah menjadi representasi *bag of words* menggunakan `dictionary.doc2bow()`, di mana setiap dokumen direpresentasikan sebagai *list* tuple (id_kata, frekuensi_kata). Implementasinya dapat dilihat pada Gambar 4.10 pada baris kesembilan. Model TF-IDF kemudian dibangun menggunakan `TfidfModel()`, yang menghitung bobot TF-IDF untuk setiap kata dalam *corpus* dan hasilnya disimpan dalam `corpus_tfidf`. Implementasinya dapat dilihat pada Gambar 4.10 di baris ke-12 hingga 13.

4.5 *Topic Modeling* dengan LDA

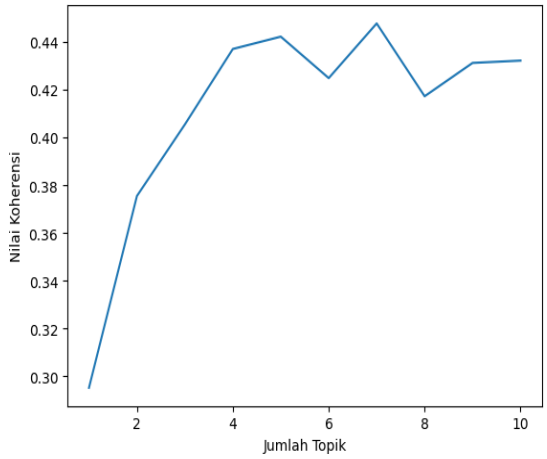
Data yang telah dilakukan pembobotan selanjutnya dilakukan analisis dengan *topic modeling* untuk mengidentifikasi topik-topik yang dibicarakan pada suatu dokumen atau data yang telah dikumpulkan. Metode *Latent Dirichlet Allocation* (LDA) diterapkan pada penelitian ini untuk *topic modeling* dengan menggunakan *coherence score* dalam menentukan banyaknya model. Semakin tinggi *coherence score*, semakin baik juga pemahaman interpretasi manusia terhadap topik tersebut. Tahap evaluasi model topik dilakukan dengan menghitung *coherence score* menggunakan modul `CoherenceModel` pada *library* Gensim. Pada setiap iterasi, model LDA dilatih menggunakan `LdaModel` dengan *corpus*, *dictionary*, num_topics, *random state*, jumlah *iterations,* dan jumlah *passes* yang telah ditentukan untuk mendapatkan topik dengan model yang optimal. Implementasi menghitung *coherence score* terdapat pada Gambar 4.11.

```
(1)    def compute_coherence_values(dictionary, corpus, texts, limit,
(2)    start=1, step=1):
(3)        coherence_values = []
(4)        model_list = []
(5)
(6)        for num_topics in range(start, limit, step):
(7)            model = LdaModel(
(8)                corpus=corpus,id2word=dictionary,num_topics=num_topics,
(9)
```

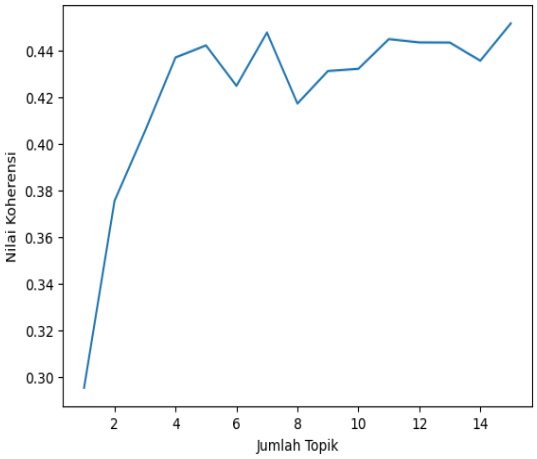
```
(10)                random_state=0, iterations=100,passes=20)
(11)
(12)            model_list.append(model)
(13)            coherencemodel = CoherenceModel(
(14)                model=model, texts=texts, dictionary=dictionary,
(15)                coherence='c_v')
(16)            coherence_values.append(coherencemodel.get_coherence())
(17)        return model_list, coherence_values
(18)
(19)    model_list, coherence_values = compute_coherence_values(
(20)        dictionary=dictionary, corpus=corpus_tfidf, texts=data,
(21)        start=1, limit=limit, step=1)
(22)
(23)    # Visualisasi hasil
(24)    x = range(1, limit)
(25)    plt.plot(x, coherence_values)
(26)    plt.xlabel("Jumlah Topik")
(27)    plt.ylabel("Nilai Koherensi")
        plt.show()
```

Gambar 4.11 Kode Program Menghitung *Coherence Score*

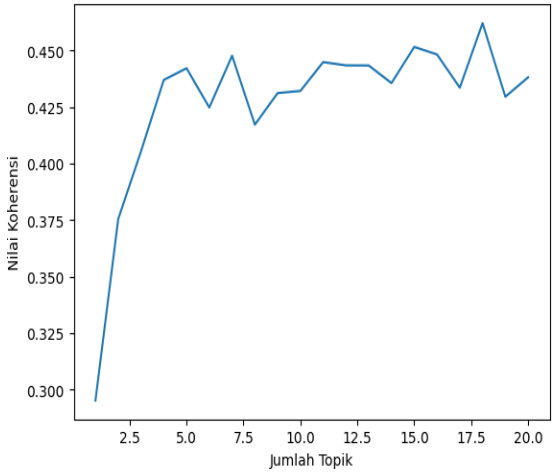
Untuk memperoleh jumlah topik yang paling optimal, dilakukan tiga kali percobaan dengan batas maksimum topik (*limit*) yang berbeda, yaitu 10, 15, dan 20. Pada masing-masing percobaan, model dilatih untuk berbagai jumlah topik mulai dari 1 hingga batas maksimum tersebut, dan nilai koherensi dari setiap model dicatat untuk dianalisis lebih lanjut. Model yang dilatih kemudian disimpan dalam `model_list` (Gambar 4.11 pada baris keempat). Selanjutnya dilakukan visualisasi berupa grafik menggunakan *library* matplotlib pada baris ke-23 hingga 27 pada Gambar 4.11, di mana sumbu x merepresentasikan jumlah topik dan sumbu y merepresentasikan nilai koherensi. Grafik dari masing-masing percobaan ditampilkan pada Gambar 4.12 hingga Gambar 4.14.



Gambar 4.12 Grafik Limit 10



Gambar 4.13 Grafik Limit 15



Gambar 4.14 Grafik Limit 20

Berdasarkan hasil evaluasi yang dirangkum dalam Tabel 4.14, setiap percobaan menghasilkan jumlah topik terbaik yang berbeda-beda. Pada percobaan pertama dengan limit 10, jumlah topik terbaik adalah 7 topik dengan nilai koherensi 0,447733. Pada percobaan kedua dengan limit 15, jumlah topik optimal adalah 15 topik dengan nilai koherensi 0,451644. Sedangkan pada percobaan ketiga dengan limit 30, jumlah topik terbaik adalah 18 topik dengan nilai koherensi 0,462177.

Tabel 4.14 *Coherence Value* (Bagian 1)

Percobaan ke	**1**		**2**		**3**	
Limit	**10**		**15**		**20**	
	Num_topic	**Coherence Value**	**Num_topic**	**Coherence Value**	**Num_topic**	**Coherence Value**
	1	0,295200	1	0,295200	1	0,295200
	2	0,375491	2	0,375491	2	0,375491
	3	0,405443	3	0,405443	3	0,405443
	4	0,437036	4	0,437036	4	0,437036
	5	0,442204	5	0,442204	5	0,442204
	6	0,424842	6	0,424842	6	0,424842
	7	**0,447733**	7	0,447733	7	0,447733
	8	0,417244	8	0,417244	8	0,417244
	9	0,431204	9	0,431204	9	0,431204
	10	0,432168	10	0,432168	10	0,432168
			11	0,444925	11	0,444925
			12	0,443479	12	0,443479
			13	0,443430	13	0,443430
			14	0,435636	14	0,435636
			15	**0,451644**	15	0,451644

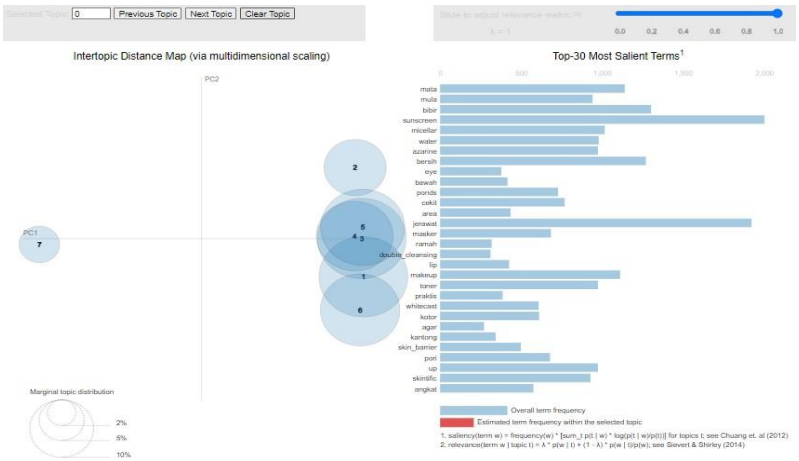
Tabel 4.14 *Coherence Value* (Bagian 2)

Percobaan ke	**1**		**2**		**3**	
Limit	**10**		**15**		**20**	
	Num_topic	**Coherence Value**	**Num_topic**	**Coherence Value**	**Num_topic**	**Coherence Value**
					16	0,448313
					17	0,433604
					18	**0,462177**
					19	0,429588
					20	0,438189
Topik optimal	**7**		**15**		**18**	

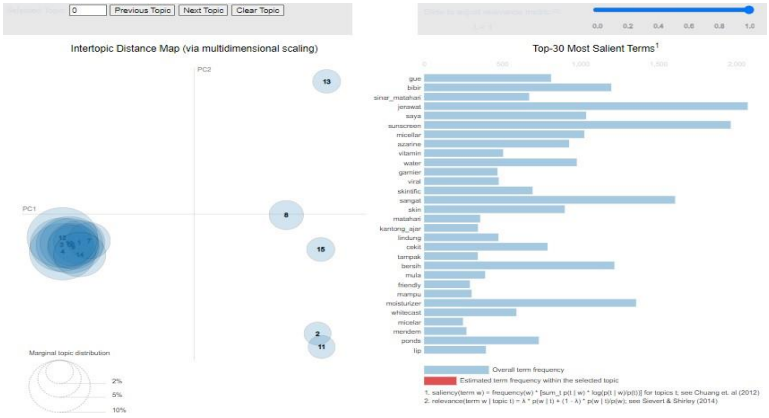
Untuk mengetahui distribusi kata antar topik, visualisasi diimplementasikan menggunakan *library* `pyLDAvis` yang dapat dilihat pada Gambar 4.15. Visualisasi ini membantu untuk melihat kata-kata yang paling relevan untuk setiap topik dan memahami hubungan antar topik. Hasil visualisasi LDA dari ketiga jumlah topik tertinggi dari masing-masing *limit* dapat dilihat pada Gambar 4.16 hingga Gambar 4.18.

```
highest_coherence_topic = optimal_model
pyLDAvis.enable_notebook()
vis = pyLDAvis.gensim.prepare(optimal_model, corpus_tfidf,dictionary,
      sort_topics=False)
pyLDAvis.display(vis)
```

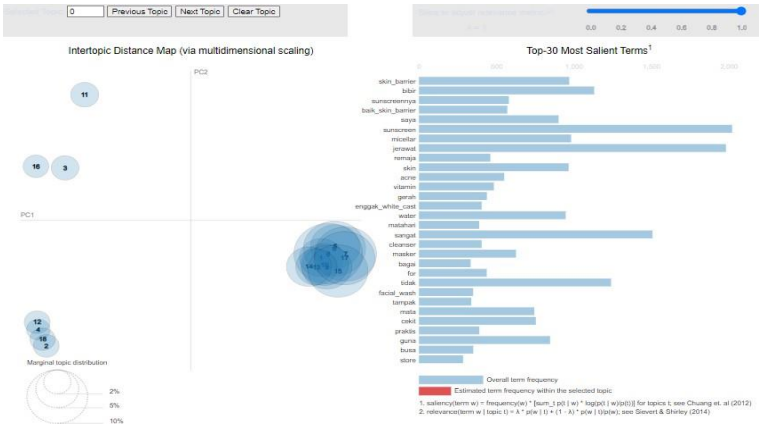
Gambar 4.15 Kode Program Visualisasi Model LDA



Gambar 4.16 Visualisasi LDA dengan Limit 10



Gambar 4.17 Visualisasi LDA dengan Limit 15



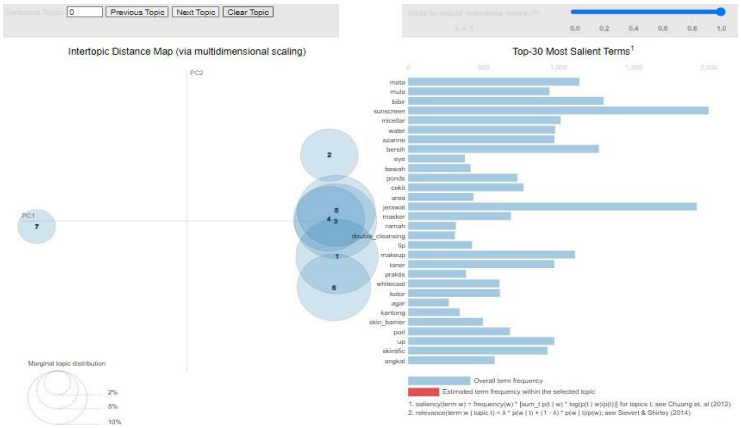
Gambar 4.18 Visualisasi LDA dengan Limit 20

Pada Gambar 4.16, terlihat bahwa topik-topik yang dihasilkan sebagian besar terkonsentrasi pada satu area di kuadran 1 dan 4, menandakan adanya kedekatan semantik antar topik. Meskipun demikian, karena jumlah topik masih relatif sedikit, tumpang tindih antar gelembung tidak terlalu padat sehingga identifikasi masing-masing topik masih memungkinkan dilakukan. Sementara itu, visualisasi pada Gambar 4.17, menunjukkan peningkatan jumlah gelembung, namun persebarannya tidak diikuti oleh peningkatan ruang visual. Hampir seluruh topik saling bertumpuk dalam satu area yang padat pada kuadran 3. Hal ini menyebabkan kesulitan dalam membedakan antar topik karena banyaknya tumpang tindih baik dari segi posisi maupun ukuran gelembung.

Visualisasi pada Gambar 4.18, menunjukkan persebaran topik yang tidak jauh berbeda dengan Gambar 4.17. Jumlah topik mengalami peningkatan, namun topik-topik Sebagian besar juga saling bertindihan pada satu area yang padat pada kuadran 4. Hal ini juga dapat menyebabkan kesulitan dalam membedakan antar topik karena banyaknya tumpang tindih baik

dari segi posisi maupun ukuran gelembung. Hal ini menunjukkan bahwa jumlah topik yang terlalu banyak justru dapat menurunkan kualitas interpretasi topik secara keseluruhan.

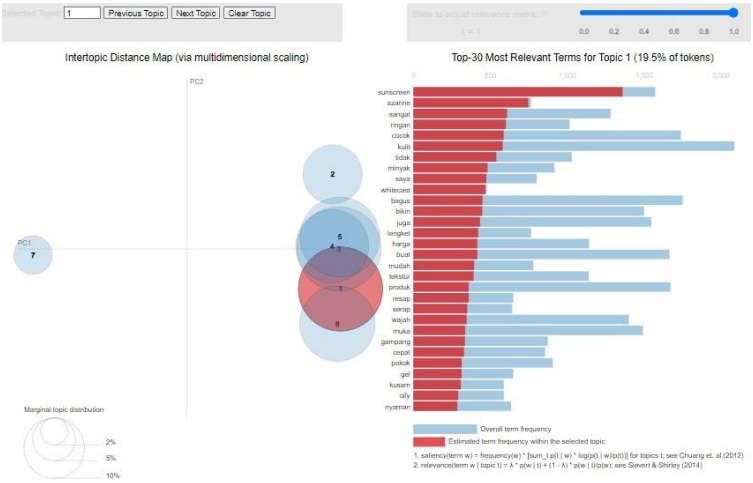
Berdasarkan ketiga visualisasi tersebut, dapat disimpulkan bahwa model dengan limit 10 menghasilkan visualisasi yang paling optimal untuk interpretasi lanjutan. Meskipun masih terdapat tumpang tindih antar beberapa topik, namun jarak antar gelembung masih cukup untuk membedakan entitas topik yang berbeda. Oleh karena itu, limit 10 dipilih sebagai jumlah topik yang digunakan dalam tahap interpretasi dan pembangunan ontologi, karena mampu menjaga keseimbangan antara koherensi semantik dan keterbacaan visual.



Gambar 4.19 Visualisasi LDA dengan Limit 10

Barplot yang terlihat pada Gambar 4.19 menampilkan frekuensi 30 kata yang sering muncul beserta visualisasi topik berbentuk gelembung. Ukuran gelembung merepresentasikan frekuensi kemunculan topik, semakin besar gelembung maka semakin tinggi frekuensi kemunculannya. Posisi gelembung yang saling berdekatan menandakan adanya keterkaitan, sementara yang berjauhan menunjukkan sebaliknya. Gelembung yang tersebar dan tidak beririsan menandakan tidak adanya keterkaitan antar topik. Visualisasi LDA pada Gambar 4.19 terdapat gelembung yang bertabrakan sehingga dapat disimpulkan jika adanya topik yang beririsan, topik yang beririsan yaitu:

- Topik 1 beririsan dengan Topik 3, 4, 5, dan 6.
- Topik 2 beririsan dengan Topik 5.
- Topik 3 beririsan dengan Topik 1, 4, 5, dan 6.
- Topik 4 beririsan dengan Topik 1, 3, dan 5.
- Topik 5 beririsan dengan Topik 1, 2, 3, dan 4.
- Topik 6 beririsan dengan Topik 1 dan 3.
- Topik 7 tidak beririsan dengan topik lain.

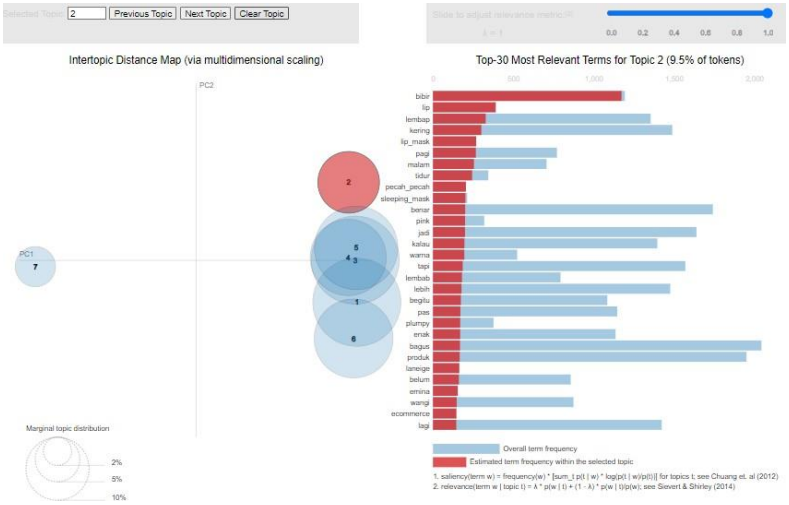


Gambar 4.20 Visualisasi *topic*-1

Gambar 4.20 menampilkan kata-kata yang muncul dalam topik ke-1. Kata-kata dengan probabilitas tertinggi yang mewakili topik ke-1 yaitu: ‘0,030*"sunscreen" + 0,016*"azarine" + 0,013*"sangat" + 0,013*"ringan" + 0,013*"cocok" + 0,013*"kulit" + 0,012*"tidak" + 0,011*"minyak" + 0,010*"saya" + 0,010*"whitecast"’. Untuk memperkuat interpretasi ini, Tabel 4.15 menyajikan beberapa kutipan komentar pengguna yang dikategorikan ke dalam topik ini berdasarkan distribusi dominannya yang menunjukkan relevansi semantik yang tinggi dengan kata-kata utama dalam topik.

Tabel 4.15 Contoh Komentar yang Relevan dengan Topik 1

Komentar	Kata Kunci yang Muncul
“sunscreen azarine ringan bangetttt teksturenya terus cepet nyerep, gak peeling dan gak bikin minyakan. cocok buat semua tipe kulit mau yang berminyak maupun kering plus harganya murahhhh dpt banyakkk sangat sangat sukaaaaaaa❤️ ❤️ ❤️ 👠👠👠❤️ ❤️ ❤️”	‘sunscreen’, ‘azarine’, ‘ringan’, ‘cocok’, ‘kulit’, ‘minyak’
“Sunscreenya bagus no whitecast, cocok disegala tipe kulit. aku udah make 1tahun lebih dan cocok. Sunscreen ini sejuta umat banget, gampang menyerap juga di wajah. semoga ada kemasan yang lebih besar lagi ya. Ayo kalian coba pake azarine buat ngejaga kulit kita dari sinar matahari”	‘sunscreen’, ‘azarine’, ‘cocok’, ‘kulit’, ‘tidak’, ‘whitecast’
my sunscreen holy grail! dulu sempet bingung cari sunscreen karena emang kulitku berminyak banget. setelah ketemu azarine bener-bener langsung fall in love! teksturnya gel dan ada sensasi coolingnya, ga whitecast, dan pastinya harga terjangkau!	‘sunscreen’, ‘azarine’, ‘kulit’, ‘tidak’, ‘minyak’, ‘whitecast’

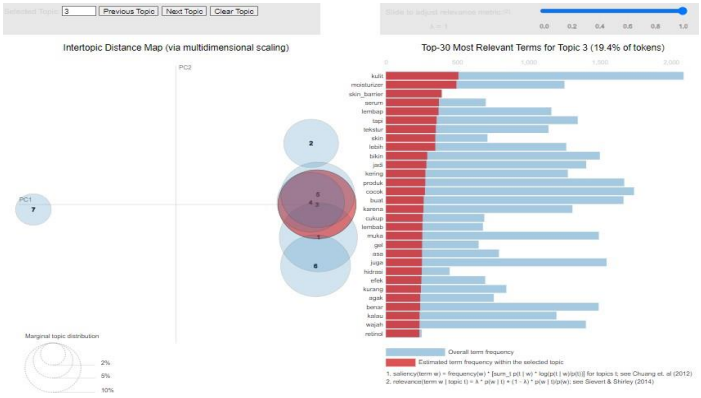


Gambar 4.21 Visualisasi *topic*-2

Gambar 4.21 menampilkan kata-kata yang muncul dalam topik ke-2. Kata-kata dengan probabilitas tertinggi yang mewakili topik ke-2 yaitu: ‘0,045*"bibir" + 0,015*"lip" + 0,013*"lembap" + 0,012*"kering" + 0,010*"lip_mask" + 0,010*"pagi" + 0,010*"malam" + 0,009*"tidur" + 0,008*"pecah_pecah" + 0,008*"sleeping_mask"’. Untuk memperkuat interpretasi ini, Tabel 4.16 menyajikan beberapa kutipan komentar pengguna yang dikategorikan ke dalam topik ini berdasarkan distribusi dominannya yang menunjukkan relevansi semantik yang tinggi dengan kata-kata utama dalam topik.

Tabel 4.16 Contoh Komentar yang Relevan dengan Topik 2

Komentar	**Kata Kunci yang Muncul**
“lip mask ter enak menurut aku, teksturnya creamy dan agak berminyak tapi tetep nyaman, karna bikin bibir lembab dan plumpy, harganya juga affordable, aku pake ini malam hari sebelum tidur, pakenya tipis” aja udah lembab, baunya juga enak kaya bau permen karet, next mau coba varian yg lain”	‘lip_mask’, ‘lembap’, ‘malam’
“Aku udah pake lip mask ini kaya hampir setahun gitu dan udah repurchase!!! kalo untuk aku, lip mask emina bikin bibir lembab dan lembut banget. tadinya bibirku pecah2 dan kadang suka sampe berdarah. ini cocok banget sih untuk yg kondisi bibirnya kaya aku.”	‘lip_mask’, ‘lembap’, ‘pecah_pecah’
“rasanya seperti risih dan lengket gitu, tapi karena dipakai tidur jadi no prob. tiap bangun tidur ngerasa enakeun banget bikin bibir lembab ga pecah-pecah, terus kayak keangkat kulit mati si bibir. produk ini habisnya lama bangettt, sampai bosan mau ngehabisinnya🙌🙌”	‘bibir’, ‘lembap’, ‘pecah_pecah’

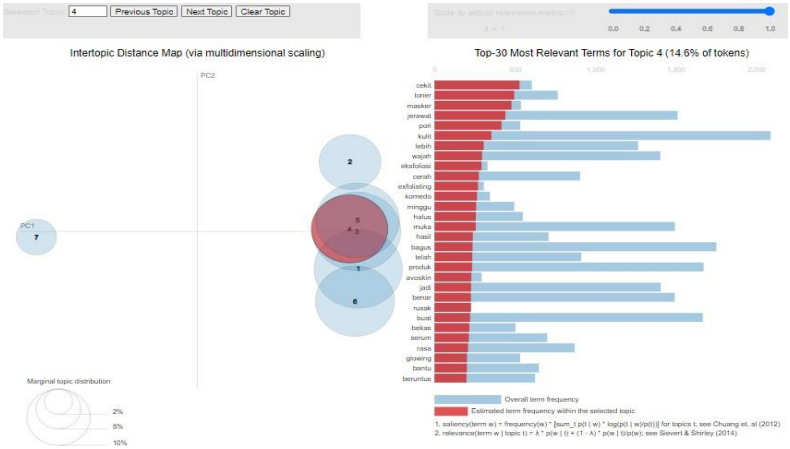


Gambar 4.22 Visualisasi *topic*-3

Gambar 4.22 menampilkan kata-kata yang muncul dalam topik ke-3. Kata-kata dengan probabilitas tertinggi yang mewakili *topic* ke-3 yaitu: ‘0,011*"kulit" + 0,011*"moisturizer" + 0,009*"skin_barrier" + 0,008*"serum" + 0,008*"lembap" + 0,008*"tapi" + 0,008*"tekstur" + 0,008*"lebih" + 0,008*"skin" + 0,006*"bikin"’. Untuk memperkuat interpretasi ini, Tabel 4.17 menyajikan beberapa kutipan komentar pengguna yang dikategorikan ke dalam topik ini berdasarkan distribusi dominannya yang menunjukkan relevansi semantik yang tinggi dengan kata-kata utama dalam topik.

Tabel 4.17 Contoh Komentar yang Relevan dengan Topik 3

Komentar	Kata Kunci yang Muncul
“Udah pake moisturizer skin barrier dear me dari zaman yang tube sampe packaging yang sekarang. Tetep setia karna sebagus & secocok itu sama aku. Teksturnya pas ga terlalu cair & gabikin lengket. Bantu banget ilangin bruntusan & bekas jerawat di kulit akuu. Luv dear me ❓ ”	‘kulit’, ‘moisturizer’, ‘skin barrier’, ‘tekstur’, ‘bikin’
“Jujur ini moisturizer paling bagus sih menurut aku, dia tuh ga lengket dikulit dan cepet banget nyerapnya jadi ga yang harus nunggu lama, terus beneran dia ngejaga banget skin barrier aku pokoknya bagus banget sih, kalau mutihin dia enggak cuman agak nyerahin gitu jadi ga bikin kusam”	‘kulit’, ‘moisturizer’, ‘skin barrier’
“Secintaaa itu sma moisturizer yg satu ini, udah coba beberapa moist tapi ini yg paling bestt, dari segi tekstur dia gel yg mudah nyerap, wrnya putih susu, pas di pakai di wajah sangat ngelembapin, gak ada aroma yg aneh", memperbaiki skin barrier, mencerahkan jugaa ❓ minusnya (-) cmn di harganya yg tidak ramah di kantong mahasiswa hihihi dgn uk 30 gr cmn bisa smpe -+ 2 mingguan ❐”	‘moisturizer’, ‘skin barrier’, ‘tapi’, ‘tekstur’

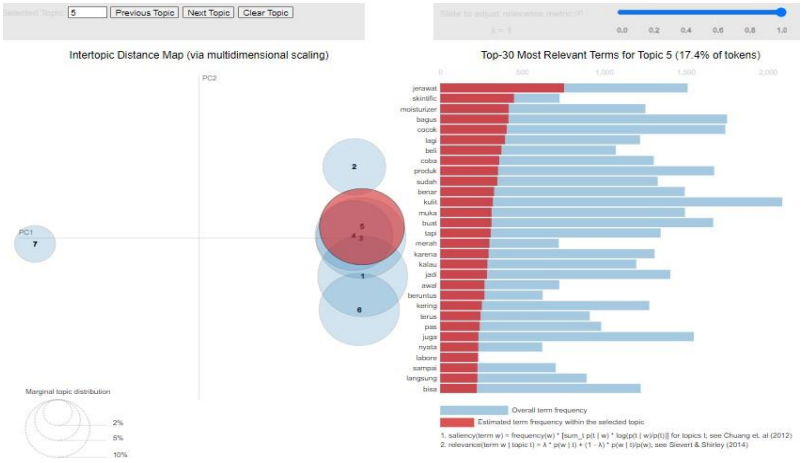


Gambar 4.23 Visualisasi *topic*-4

Gambar 4.23 menampilkan kata-kata yang muncul dalam topik ke-4. Kata-kata dengan probabilitas tertinggi yang mewakili topik ke-4 yaitu: ‘0,016*"pori" + 0,015*"toner" + 0,014*"masker" + 0,013*"jerawat" + 0,012*"pori" + 0,010*"kulit" + 0,009*"lebih" + 0,009*"wajah" + 0,009*"eksfoliasi" + 0,008*"cerah"’. Untuk memperkuat interpretasi ini, Tabel 4.18 menyajikan beberapa kutipan komentar pengguna yang dikategorikan ke dalam topik ini berdasarkan distribusi dominannya. Komentar-komentar ini menunjukkan relevansi semantik yang tinggi dengan kata-kata utama dalam topik.

Tabel 4.18 Contoh Komentar yang Relevan dengan Topik 4

Komentar	Kata Kunci yang Muncul
“direkomen dari temen buat texture & pores, baunya typical chemical exfoliator pada umumnya, gak bikin cekat cekit karena persen aha-bha nya juga gak tinggi, pernah gak sengaja pake 2 hari berturut-turut & untung gak bikin kulit iritasi sih”	‘pori’, ‘toner’, ‘kulit’, ‘bikin’
“Toner exfo pertamaku setelah sebelumnya exfo pakai clay mask. Termasuk exfo mild ya jadi nggak ada purging sama sekali waktu mulai pakai. Menurutku dia nggak cekit cekit waktu kulit lagi normal, tapi kalau ada jerawat baru kerasa lumayan cekit-cekit. Untuk matengin jerawat dia bagus bangettt, ngilangin komedo juga jadi nggak usah repot-repot pake pore strip lagi”	‘pori’, ‘toner’, ‘eksfoliasi’, ‘masker’, ‘ jerawat’, ‘kulit’
Eksfoliasi toner yang cocok di semua jenis kulit yang mengandung aha bha, untuk teksturnya sendiri cair, digunakan saat malam hari 2-3 kali dalam seminggu, after pemakaian paginya wajah cerah dan kulit terasa halus, mengangkat sel kulit mati dan bisa mengecilkan pori pori	‘eksfoliasi’, ‘toner’, ‘kulit’, ‘wajah’, ‘cerah’, ‘pori’

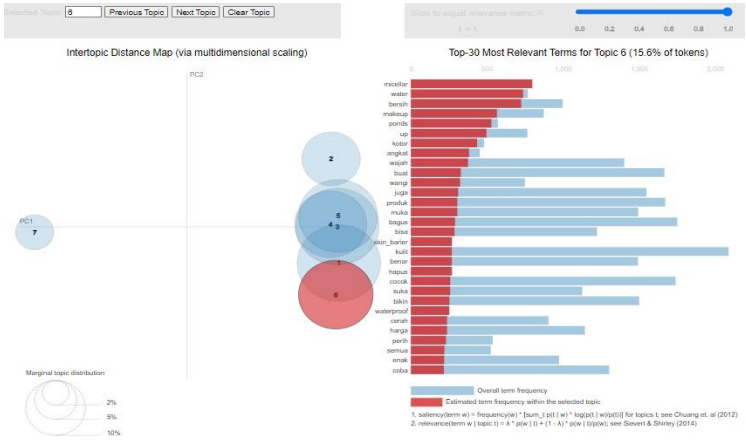


Gambar 4.24 Visualisasi *topic*-5

Gambar 4.24 menampilkan kata-kata yang muncul dalam topik ke-5. Kata-kata dengan probabilitas tertinggi yang mewakili topik ke-5 yaitu: ‘0,019*"jerawat" + 0,011*"skintific" + 0,010*"moisturizer" + 0,010*"bagus" + 0,010*"cocok" + 0,010*"lagi" + 0,009*"beli" + 0,009*"coba" + 0,009*"produk" + 0,009*"sudah"’. Untuk memperkuat interpretasi ini, Tabel 4.19 menyajikan beberapa kutipan komentar pengguna yang dikategorikan ke dalam topik ini berdasarkan distribusi dominannya. Komentar-komentar ini menunjukkan relevansi semantik yang tinggi dengan kata-kata utama dalam topik.

Tabel 4.19 Contoh Komentar yang Relevan dengan Topik 5

Komentar	Kata Kunci yang Muncul
“Pertama kali tau moiturizer ini dari temen dan ternyata dia cocok pakai ini karena sebelumnya kulit skin bariernya sangat kusam dan setelah pakai skintific yg moisturizer skin bariernya kembali bersih makanya aku coba pakai dan hampir setahun aku pakai skintific ini”	'moisturizer', 'cocok', 'skintific', 'coba', 'sudah'
“skintific moisturizer ini bagus bangett, udah beli ke 3x nya yg ukuran 80g ngelembabin banget, pokoknya tiap udah pake ini kulit terasa segar, kenyal, dan lembab, worth it baget!! semenjak pake ini juga jadi jarang jerawatan, pokoknya kalian harus cobainn”	'skintific', 'moisturizer', 'bagus', 'beli', 'sudah', 'jerawat', 'coba'
“pengalaman pertama pake skintific moisturizer tuh ringan banget, cocok banget buat muka ku walopun banyak jerawat nya, next time aq bakal pilih lagi deh brand skintific nya, semoga bikin wajahku tambah cerah luph you skintific moisturizer”	'skintific', 'moisturizer', 'cocok', 'jerawat', 'lagi

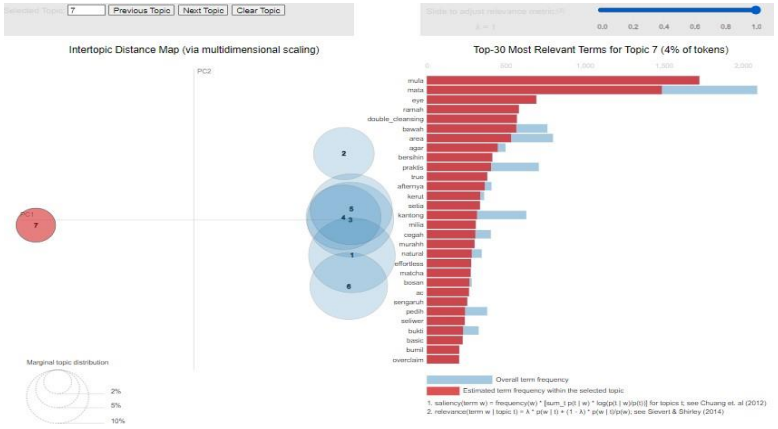


Gambar 4.25 Visualisasi *topic*-6

Gambar 4.25 menampilkan kata-kata yang muncul dalam topik ke-6. Kata-kata dengan probabilitas tertinggi yang mewakili topik ke-6 yaitu: ‘0,022*"micellar" + 0,020*"water" + 0,020*"bersih" + 0,016*"makeup" + 0,015*"ponds" + 0,014*"up" + 0,012*"kotor" + 0,011*"angkat" + 0,010*"wajah" + 0,009*"buat"’. Untuk memperkuat interpretasi ini, Tabel 4.20 menyajikan beberapa kutipan komentar pengguna yang dikategorikan ke dalam topik ini berdasarkan distribusi dominannya. Komentar-komentar ini menunjukkan relevansi semantik yang tinggi dengan kata-kata utama dalam topik.

Tabel 4.20 Contoh Komentar yang Relevan dengan Topik 6

Komentar	Kata Kunci yang Muncul
“Micellar water pond's baguss banget buat bersihin wajah dari make up seharian.. selain harganya murah kemasannya juga praktis banget buat dibawa kemana-mana dan aku sering pake micellar water ini sehabis aktivitas juga untuk ngangkat debu di wajah aku”	'micellar', 'water', 'ponds', 'buat', 'wajah', 'makeup', 'angkat'
“ENAK BANGETT bener2 sensainya lembutt, Ponds Micellar Water oke banget sih buat bersihin muka. Makeup langsung keangkat tanpa perlu gosok keras, plus rasanya ringan di kulit. Gak bikin lengket atau kering, jadi enak dipake sehari-hari. Simple tapi manjur, cocok buat yang suka skincare santai!”	'micellar', 'water', 'ponds', 'buat', 'makeup', 'angkat'
“suka banget sama micellar water ponds ini. Dia ini mengandung 5 bahan aktif yang bisa mencerahkan wajah salah satunya niacinamide. Micellar water ini juga ringan banget untuk kulit sensitif aku ini dan cukup ampuh untuk bersihin make up tebel dan yg paling penting dia ini non alcohol luv bngt deh”	'micellar', 'water', 'ponds', 'wajah', 'makeup', 'bersih'



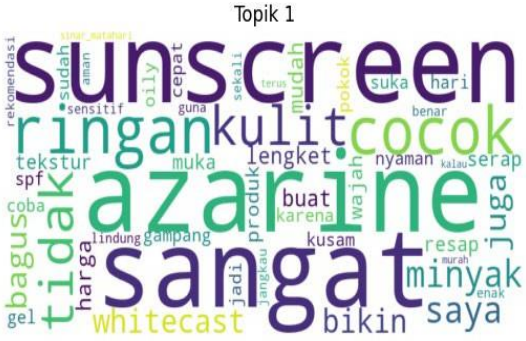
Gambar 4.26 Visualisasi *topic*-7

Gambar 4.26 menampilkan kata-kata yang muncul dalam topik ke-7. Kata-kata dengan probabilitas tertinggi yang mewakili topik ke-7 yaitu: ‘0,079*"mula" + 0,068*"mata" + 0,032*"eye" + 0,027*"ramah" + 0,026*"double_cleansing" + 0,026*"bawah" + 0,024*"area" + 0,020*"agar" + 0,019*"bersihin" + 0,019*"praktis"’. Untuk memperkuat interpretasi ini, Tabel 4.21 menyajikan beberapa kutipan komentar pengguna yang dikategorikan ke dalam topik ini berdasarkan distribusi dominannya. Komentar-komentar ini menunjukkan relevansi semantik yang tinggi dengan kata-kata utama dalam topik.

Tabel 4.21 Contoh Komentar yang Relevan dengan Topik 7

Komentar	Kata Kunci yang Muncul
“Enakeun dipake buat bersihin wajah ga ninggalin rasa berminyak diwajah/ malah bikin kering. Bersihnya berasa. Diarea mata ga bikin pedih/ panas. Bikin kulit tetep lembab. Mau repurchase gampang cari dan harganya affordable.”	'bersihin', 'mata', 'area'
“Please bgt untuk yang suka pakai Garnier Micellar Water, wajib cobain yang rose ini ? □□ seenak itu, wangi, ringan, mantul lebih bagus daripada yang tutup pink / biru biasanya itu. Memang tidak se oily yang micellar water kuning yg ada oil based nya, tapi pakai ini bener2 bisa ampuh juga untuk ngebersihin area mata yang full eyeshadow bold ? ”	'bersihin', 'area', 'mata'
“gampang buat hapus eyeshadow, lem bulu mata, mascara, dan eyebrow pencil. Pakenya juga irit banget gausah banyak banyak, tapi kalo aku kapasnya di tempel ke kelopak mata agak didiemin sambil di press biasa aja pake tangan terus baru dibersihin, hasilnya udh lumayan bersih, tinggal double cleansing pake facial wash.”	'mata', 'double_cleansing', 'bersihin'

Visualisasi selanjutnya dengan menggunakan *wordcloud*. Visualisasi *wordcloud* digunakan untuk mempermudah memperoleh wawasan dari topik-topik yang dihasilkan oleh pemodelan (Naury et al., 2021). Visualisasi ini berdasarkan hasil pemodelan LDA pada Gambar 4.20 hingga Gambar 4.26, sehingga terdapat 7 visualisasi *wordcloud* yang menunjukkan hasil pemodelan topik. Distribusi kata tiap topik di mana besar kecilnya kata mempunyai arti terhadap tingkat kemunculan kata. Adapun visualisasi *wordcloud* dapat dilihat pada Gambar 4.27 sampai Gambar 4.33. Implementasi kode untuk visualisasi *wordcloud* untuk tiap *topic* dari model optimal ditunjukkan pada Gambar 4.34.



Gambar 4.27 *Worldcloud 1*



Gambar 4.28 *Worldcloud 2*



Gambar 4.29 *Worldcloud 3*



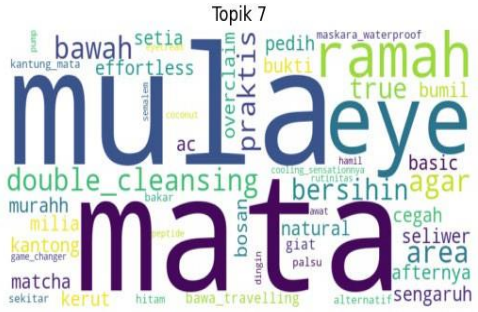
Gambar 4.30 *Worldcloud 4*



Gambar 4.31 *Worldcloud 5*



Gambar 4.32 *Worldcloud 6*



Gambar 4.33 *Worldcloud 7*

```
# Visualisasi word cloud untuk setiap topik dari model optimal
for i in range(optimal_model.num_topics):
    temp = optimal_model.show_topic(i, topn=50)
    topic_words = " ".join([word for word, prop in temp])

   #Generate wordcloud
      wordcloud=WordCloud(width=800,height=400,background_color='white').ge
nerate(topic_words)
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis("off")
    plt.title("Topik " + str(i + 1))
    plt.show()
```

Gambar 4.34 Implementasi *Worldcloud*

Implementasi kode pada Gambar 4.34 menunjukkan proses pembuatan *wordcloud* untuk masing-masing topik dari model LDA yang telah dipilih sebagai model optimal berdasarkan nilai *coherence score*. Pada bagian awal perulangan, dilakukan iterasi terhadap seluruh topik dengan `range(optimal_model.num_topics)` yang secara dinamis menyesuaikan jumlah topik dalam model. Untuk setiap iterasi `i`, fungsi `optimal_model.show_topic(i, topn=50)` digunakan untuk mengekstrak 50 kata teratas dari topik ke-`i`, beserta nilai proporsi kontribusinya terhadap topik tersebut. Kemudian, hanya kata-kata dari pasangan `(word, prop)` yang diambil dan disatukan menggunakan metode `" ".join()` untuk membentuk satu string panjang `topic_words`.

Selanjutnya, objek *wordcloud* dibentuk menggunakan fungsi `WordCloud()` dari pustaka *wordcloud*, dengan parameter `width=800`, `height=400`, dan `background_color='white'` untuk menghasilkan tampilan visual yang bersih dan jelas. Fungsi `.generate(topic_words)` diterapkan untuk membangun *word cloud* berdasarkan kumpulan kata yang telah dihimpun sebelumnya. Visualisasi ditampilkan menggunakan `matplotlib.pyplot` (`plt`) dengan fungsi `imshow()` untuk menampilkan

gambar, `axis("off")` untuk menyembunyikan sumbu koordinat, dan `title()` untuk memberi judul sesuai nomor topik. Proses ini diulang untuk setiap topik sehingga menghasilkan *wordcloud* yang merepresentasikan kata-kata dominan pada setiap topik.

4.6 Hasil dan Analisis

Berdasarkan hasil *topic modeling*, maka penulis akan menganalisis secara kualitatif. Hasil dari *topic modeling* akan ditafsirkan makna yang terkandung dalam kata-kata yang muncul di tiap topik. Penafsiran makna dari suatu topik diambil berdasarkan tujuh kata yang paling dominan seperti yang ditunjukan pada Tabel 4.22.

Tabel 4.22 Analisis Hasil *Topic Modeling*

Topic ke-	Model
1	‘0,030*"sunscreen" + 0,016*"azarine" + 0,013*"sangat" + 0,013*"ringan" + 0,013*"cocok" + 0,013*"kulit" + 0,012*"tidak" + 0,011*"minyak" + 0,010*"saya" + 0,010*"whitecast"’
2	‘0,045*"bibir" + 0,015*"lip" + 0,013*"lembap" + 0,012*"kering" + 0,010*"lip_mask" + 0,010*"pagi" + 0,010*"malam" + 0,009*"tidur" + 0,008*"pecah_pecah" + 0,008*"sleeping_mask"’
3	‘0,011*"kulit" + 0,011*"moisturizer" + 0,009*"skin_barrier" + 0,008*"serum" + 0,008*"lembap" + 0,008*"tapi" + 0,008*"tekstur" + 0,008*"lebih" + 0,008*"skin" + 0,006*"bikin"’
4	‘0,016*"cekit" + 0,015*"toner" + 0,014*"masker" + 0,013*"jerawat" + 0,012*"pori" + 0,010*"kulit" + 0,009*"lebih" + 0,009*"wajah" + 0,009*"eksfoliasi" + 0,008*"cerah"’
5	‘0,019*"jerawat" + 0,011*"skintific" + 0,010*"moisturizer" + 0,010*"bagus" + 0,010*"cocok" + 0,010*"lagi" + 0,009*"beli" + 0,009*"coba" + 0,009*"produk" + 0,009*"sudah"’
6	‘0,022*"micellar" + 0,020*"water" + 0,020*"bersih" + 0,016*"makeup" + 0,015*"ponds" + 0,014*"up" + 0,012*"kotor" + 0,011*"angkat" + 0,010*"wajah" + 0,009*"buat"’
7	‘0,079*"mula" + 0,068*"mata" + 0,032*"eye" + 0,027*"ramah" + 0,026*"double_cleansing" + 0,026*"bawah" + 0,024*"area" + 0,020*"agar" + 0,019*"bersihin" + 0,019*"praktis"’

Topik 1 menampilkan kata kunci seperti *sunscreen*, *azarine*, ringan, cocok, kulit, *whitecast*, dan minyak. Topik ini menunjukkan bahwa diskusi berfokus pada pengalaman pengguna dalam menggunakan produk *sunscreen*, khususnya merek Azarine, yang dinilai ringan, cocok untuk berbagai jenis kulit, dan tidak menimbulkan *whitecast* atau produksi minyak berlebih. Fokus utama dari topik ini terletak pada karakteristik tekstur dan efektivitas perlindungan terhadap sinar matahari.

Topik 2 didominasi oleh kata-kata seperti bibir, *lip*, lembap, kering, *lip_mask*, malam, pagi, dan *sleeping_mask*, yang menggambarkan diskusi terkait perawatan bibir, terutama penggunaan *lip mask*. Produk-produk yang dimaksud kemungkinan digunakan sebagai perawatan intensif di malam hari untuk mengatasi bibir pecah-pecah atau kehilangan kelembapan.

Topik 3 mengandung kata-kata seperti *moisturizer*, kulit, *skin_barrier*, serum, tekstur, lembap, dan bikin. Topik ini mencerminkan pembahasan mengenai pelembap dan serum yang dikaitkan dengan kemampuan memperkuat *skin barrier* dan memperbaiki tekstur kulit. Kehadiran kata seperti “bikin” dan “tapi” menunjukkan adanya evaluasi subjektif dari pengguna dalam menilai efektivitas suatu produk terhadap kondisi kulit mereka.

Topik 4 mencakup kata seperti cekit, *toner*, *masker*, jerawat, pori, eksfoliasi, dan cerah, yang menunjukkan pembahasan mengenai produk-produk eksfoliasi atau pembersih seperti *toner* dan *masker*, khususnya yang memiliki efek sensasi perih (cekit) saat digunakan. Produk-produk ini ditujukan untuk membersihkan pori-pori, membantu mencerahkan kulit, serta mengatasi jerawat melalui proses eksfoliasi.

Topik 5 menyoroti isu jerawat melalui kata kunci seperti jerawat, *skintific*, *moisturizer*, cocok, lagi, produk, dan coba. Topik ini menampilkan pengalaman pengguna dalam mencoba berbagai produk perawatan jerawat, terutama dari merek Skintific. Kehadiran kata-kata seperti “cocok”, “lagi”, dan “sudah” merefleksikan evaluasi berkelanjutan terkait keefektifan produk dalam mengatasi jerawat.

Topik 6 memuat kata-kata seperti *micellar*, *water*, bersih, *makeup*, *ponds*, kotor, angkat, dan wajah, yang mengindikasikan diskusi mengenai produk pembersih wajah seperti *micellar water*. Produk ini digunakan untuk mengangkat sisa *makeup* dan kotoran secara efektif, dengan merek Pond’s sebagai salah satu yang paling banyak disebut dalam percakapan pengguna.

Topik 7 didominasi oleh kata-kata mula, mata, *eye*, *double_cleansing*, ramah, praktis, dan bersihin. Kata kunci seperti "mata", "*eye*", "area", dan "bawah" menunjukkan bahwa pengguna sering membahas pengalaman membersihkan riasan di sekitar mata, yang merupakan area sensitif. Sementara itu, kata "mula" mengindikasikan bahwa diskusi ini relevan bagi pengguna yang baru memulai rutinitas *skincare* atau teknik *double cleansing*. Kata-kata seperti "ramah", "praktis", dan "agar" mencerminkan preferensi terhadap produk yang tidak hanya efektif, tetapi juga mudah digunakan dan aman, terutama bagi pemula.

Secara umum, ketujuh topik yang dihasilkan memperlihatkan tingkat koherensi semantik yang baik dan relevan terhadap domain *skincare*. Setiap topik berhasil menangkap subtema

yang khas seperti perlindungan sinar matahari, perawatan bibir, kelembapan kulit, eksfoliasi, penanganan jerawat, pembersihan wajah, serta perawatan area mata. Tidak ditemukan topik yang tidak koheren secara semantik, sehingga masing-masing topik dapat diinterpretasikan dengan baik dan mewakili dinamika diskusi di komunitas pengguna *skincare*.

Untuk memperoleh pemahaman yang lebih komprehensif mengenai struktur semantik dari hasil pemodelan topik, dilakukan analisis terhadap irisan antar topik yang divisualisasikan menggunakan pustaka pyLDAvis. Hasil visualisasi menunjukkan bahwa sebagian besar topik yang dihasilkan saling beririsan, dengan pengecualian pada Topik 7 yang tampil terpisah dan tidak menunjukkan hubungan semantik signifikan dengan topik lainnya. Topik 1, yang berkaitan erat dengan diskusi mengenai *sunscreen*, khususnya dari merek Azarine, menunjukkan irisan kuat dengan Topik 3, 4, 5, dan 6. Hal ini mengindikasikan bahwa isu-isu seperti tekstur ringan, ketidaknyamanan seperti minyak berlebih atau *whitecast*, serta kecocokan terhadap kulit merupakan tema sentral yang muncul tidak hanya dalam konteks perlindungan sinar matahari, tetapi juga pada produk pelembap (Topik 3), penanganan jerawat (Topik 5), dan pembersih wajah seperti *micellar water* (Topik 6). Kata kunci seperti “ringan”, “kulit”, dan “cocok” menjadi penghubung semantik antar topik-topik tersebut.

Topik 2, yang berfokus pada perawatan bibir menggunakan *lip mask* atau *lip sleeping mask*, menunjukkan irisan dengan Topik 5 yang menekankan pengalaman pengguna terhadap produk jerawat dan pelembap. Keterhubungan ini memperlihatkan bahwa pengguna cenderung mengaitkan aspek kelembapan dan pemulihan kulit tidak hanya pada area wajah secara umum, tetapi juga pada area bibir yang rentan terhadap kekeringan. Irisan ini didukung oleh kemunculan kata-kata seperti “lembap”, “kering”, dan “produk”, yang mengindikasikan adanya perhatian serupa terhadap hidrasi kulit.

Topik 3 menjadi salah satu topik sentral yang menunjukkan keterkaitan semantik dengan Topik 1, 4, 5, dan 6. Fokus pada kata kunci seperti “*moisturizer*”, “*skin barrier*”, dan “tekstur” mengindikasikan bahwa isu pelembap dan perlindungan kulit menjadi tema lintas topik. Hubungan dengan Topik 4 (yang membahas toner dan eksfoliasi) memperlihatkan adanya kontinuitas rutinitas skincare, di mana pengguna cenderung menggunakan pelembap setelah eksfoliasi untuk memulihkan lapisan kulit. Sementara itu, irisan dengan Topik 6 juga logis, mengingat pembersihan wajah dengan *micellar water* sering menjadi tahap awal sebelum penggunaan serum atau pelembap.

Topik 4 yang mengangkat perbincangan mengenai toner, masker, dan eksfoliasi kulit, menunjukkan irisan semantik yang cukup luas dengan Topik 1, 3, dan 5. Tumpang tindih ini

menegaskan bahwa perawatan kulit, terutama yang berorientasi pada pengangkatan sel kulit mati dan pembersihan pori, sering kali dikaitkan dengan produk pendukung lainnya seperti sunscreen dan pelembap. Terlebih lagi, eksfoliasi kerap menjadi pemicu sensitivitas kulit, sehingga mendorong pengguna untuk mencari produk dengan fungsi menenangkan atau memperbaiki *skin barrier*.

Topik 5, yang berpusat pada permasalahan jerawat dan penggunaan produk seperti *moisturizer* dan *serum*, memiliki irisan dengan hampir seluruh topik lainnya, yaitu Topik 1, 2, 3, dan 4. Hal ini dapat dijelaskan oleh fakta bahwa jerawat merupakan salah satu permasalahan kulit paling umum dan kompleks, yang seringkali membutuhkan pendekatan multi-produk mulai dari pembersih, eksfoliator, hingga pelembap. Kata-kata seperti "cocok", "produk", dan "coba" mengindikasikan adanya eksplorasi berkelanjutan oleh pengguna dalam menemukan solusi yang efektif untuk permasalahan ini.

Topik 6, yang memuat diskusi mengenai produk pembersih wajah seperti *micellar water*, memperlihatkan irisan yang kuat dengan Topik 1 dan 3. Keterhubungan ini merefleksikan keterpaduan antara tahapan pembersihan dan tahapan perlindungan atau pelembapan kulit dalam rutinitas *skincare* harian. Aspek seperti "ringan", "bersih", dan "wajah" menjadi kata kunci yang memperkuat koneksi semantik antar topik tersebut.

Sebagai tahap tambahan untuk memperkuat interpretasi hasil pemodelan topik, dilakukan proses validasi berbasis penilaian manusia (*human judgment*). Meskipun tidak termasuk dalam tahapan utama pemodelan LDA, validasi ini bertujuan untuk meninjau kembali apakah kumpulan kata kunci dalam tiap topik benar-benar mewakili tema yang logis dan sesuai dengan konteks pembahasan di komunitas pengguna *skincare*. Validasi dilakukan oleh 2 evaluator:

1. Evaluator 1: *Beauty influencer* yang aktif dalam media sosial dengan total pengikut mencapai 4.000 di TikTok dan 2.000 di Instagram.
2. Evaluator 2: Peneliti sendiri sebagai pengguna aktif Female Daily dan *beauty enthusiast.*

Validasi dilakukan dengan mengevaluasi tujuh topik yang diperoleh dari model terbaik, yang dipilih berdasarkan nilai koherensi tertinggi. Untuk setiap topik, dianalisis kumpulan kata kunci utama yang muncul serta tiga kutipan komentar sebagai konteks. Penulis kemudian menilai keterkaitan antara kata kunci dan komentar yang relevan menggunakan skala empat tingkat, yaitu **Tidak Relevan (1), Cukup Relevan (2), Relevan (3), dan Sangat Relevan (4)**.

Sebuah topik dikategorikan "Valid" apabila mendapatkan skor "Relevan" atau lebih tinggi dalam penilaian ini. Validasi ini tidak hanya menekankan koherensi semantik antar kata kunci, tetapi juga mempertimbangkan keterwakilan makna dari komentar asli pengguna,

sehingga dapat menggambarkan realitas diskusi yang terjadi di komunitas *skincare* secara lebih akurat. Melalui pengalaman langsung penulis sebagai pengguna, validasi ini memberikan justifikasi yang kuat bahwa topik-topik yang dihasilkan memang memiliki keterkaitan tematik yang relevan dengan realitas pengguna *skincare*. Proses validasi ini merujuk pada pendekatan yang digunakan dalam studi yang dilakukan oleh Puspita dkk. (2024), yang menekankan pentingnya keterlibatan penilaian manusia dalam meninjau sejauh mana topik-topik hasil LDA yang bersifat *unsupervised* dapat dipahami dan benar-benar merepresentasikan makna yang dimaksud.

Sebagai bagian dari proses validasi topik, penulis melakukan sesi evaluasi bersama responden yang berperan sebagai evaluator 2 melalui panggilan video (*video call*) menggunakan aplikasi WhatsApp. Dokumentasi kegiatan ini disajikan pada Gambar 4.35 sebagai bukti pelaksanaan evaluasi secara langsung.



Gambar 4.35 Alur Pembuatan Ontologi

Tabel 4.23 Hasil Validasi Evaluator (Bagian 1)

Topik	Kata Kunci	Sampel Komentar	Evaluator 1	Evaluator 2
1	"sunscreen", "azarine", "sangat", "ringan", "cocok", "kulit", "tidak", "minyak", "saya", "whitecast"	“sunscreen azarine ringan bangetttt teksturenya terus cepet nyerep, gak peeling dan gak bikin minyakan. cocok buat semua tipe kulit mau yang berminyak maupun kering plus harganya murahhhh dpt banyakkk sangat sangat sukaaaaaaa❤️ ❤️ ❤️ 👏👏👏 ❤️ ❤️ ❤️”	4	4
		“Sunscreenya bagus no whitecast, cocok disegala tipe kulit. aku udah make 1tahun lebih dan cocok. Sunscreen ini sejuta umat banget, gampang menyerap juga di wajah. semoga ada kemasan yang lebih besar lagi ya. Ayo kalian coba pake azarine buat ngejaga kulit kita dari sinar matahari”	4	4

Tabel 4.23 Hasil Validasi Evaluator (Bagian 2)

Topik	Kata Kunci	Sampel Komentar	Evaluator 1	Evaluator 2
		my sunscreen holy grail! dulu sempet bingung cari sunscreen karena emang kulitku berminyak banget. setelah ketemu azarine bener-bener langsung fall in love! teksturnya gel dan ada sensasi coolingnya, ga whitecast, dan pastinya harga terjangkau!	3	4
2	"bibir", "lip", "lembap", "kering", "lip_mask", "pagi", "malam", "tidur", "pecah_pecah", "sleeping_mask"	“lip mask ter enak menurut aku, teksturnya creamy dan agak berminyak tapi tetep nyaman, karna bikin bibir lembab dan plumpy, harganya juga affordable, aku pake ini malam hari sebelum tidur, pakenya tipis” aja udah lembab, baunya juga enak kaya bau permen karet, next mau coba varian yg lain”	4	4
		“Aku udah pake lip mask ini kaya hampir setahun gitu dan udah repurchase!!! kalo untuk aku, lip mask emina bikin bibir lembab dan lembut banget. tadinya bibirku pecah2 dan kadang suka sampe berdarah. ini cocok banget sih untuk yg kondisi bibirnya kaya aku.”	4	4
		“rasanya seperti risih dan lengket gitu, tapi karena dipakai tidur jadi no prob. tiap bangun tidur ngerasa enakeun banget bikin bibir lembab ga pecah-pecah, terus kayak keangkat kulit mati si bibir. produk ini habisnya lama bangettt, sampai bosan mau ngehabisinnya😭😭”	4	4
3	"kulit", "moisturizer", "skin_barrier", "serum", "lembap", "tapi", "tekstur", "lebih", "skin", "bikin"	“Udah pake moisturizer skin barrier dear me dari zaman yang tube sampe packaging yang sekarang. Tetep setia karna sebagus & secocok itu sama aku. Teksturnya pas ga terlalu cair & gabikin lengket. Bantu banget ilangin bruntusan & bekas jerawat di kulit akuu. Luv dear me ▭”	3	3
		“Jujur ini moisturizer paling bagus sih menurut aku, dia tuh ga lengket dikulit dan cepet banget nyerapnya jadi ga yang harus nunggu lama, terus beneran dia ngejaga banget skin barrier aku pokoknya bagus banget sih, kalau mutihin dia enggak cuman agak nyerahin gitu jadi ga bikin kusam”	3	4

Tabel 4.23 Hasil Validasi Evaluator (Bagian 3)

Topik	Kata Kunci	Sampel Komentar	Evaluator 1	Evaluator 2
		“Secintaaa itu sma moisturizer yg satu ini, udah coba beberapa moist tapi ini yg paling bestt, dari segi tekstur dia gel yg mudah nyerap, wrnya putih susu, pas di pakai di wajah sangat ngelembapin, gak ada aroma yg aneh", memperbaiki skin barrier, mencerahkan jugaa □ minusnya (-) cmn d harganya yg tidak ramah di kantong mahasiswa hihihi dgn uk 30 gr cmn bisa smpe -+ 2 mingguan □”	2	3
4	"cekit", "toner", "masker", "jerawat", "pori", "kulit", "lebih", "wajah", "eksfoliasi", "cerah"	“direkomen dari temen buat texture & pores, baunya typical chemical exfoliator pada umumnya, gak bikin cekat cekit karena persen aha-bha nya juga gak tinggi, pernah gak sengaja pake 2 hari berturut-turut & untung gak bikin kulit iritasi sih”	4	3
		“Toner exfo pertamaku setelah sebelumnya exfo pakai clay mask. Termasuk exfo mild ya jadi nggak ada purging sama sekali waktu mulai pakai. Menurutku dia nggak cekit cekit waktu kulit lagi normal, tapi kalau ada jerawat baru kerasa lumayan cekit-cekit. Untuk matengin jerawat dia bagus bangettt, ngilangin komedo juga jadi nggak usah repot-repot pake pore strip lagi”	3	3
		Eksfoliasi toner yang cocok di semua jenis kulit yang mengandung aha bha, untuk teksturnya sendiri cair, digunakan saat malam hari 2-3 kali dalam seminggu, after pemakaian paginya wajah cerah dan kulit terasa halus, mengangkat sel kulit mati dan bisa mengecilkan pori pori	4	4
5	"jerawat", "skintific", "moisturizer", "bagus", "cocok", "lagi", "beli", "coba", "produk", "sudah"	“Pertama kali tau moiturizer ini dari temen dan ternyata dia cocok pakai ini karena sebelumnya kulit skin bariernya sangat kusam dan setelah pakai skintific yg moisturizer skin bariernya kembali bersih makanya aku coba pakai dan hampir setahun aku pakai skintific ini”	4	3
		“skintific moisturizer ini bagus bangett, udah beli ke 3x nya yg ukuran 80g ngelembabin banget, pokoknya tiap udah pake ini kulit terasa segar, kenyal, dan lembab, worth it baget!! semenjak pake ini juga jadi jarang jerawatan, pokoknya kalian harus cobainn”	4	4

Tabel 4.23 Hasil Validasi Evaluator (Bagian 4)

Topik	Kata Kunci	Sampel Komentar	Evaluator 1	Evaluator 2
		“pengalaman pertama pake skintific moisturizer tuh ringan banget, cocok banget buat muka ku walopun banyak jerawat nya, next time aq bakal pilih lagi deh brand skintific nya, semoga bikin wajahku tambah cerah luph you skintific moisturizer”	3	4
6	"micellar", "water", "bersih", "makeup", "ponds", "up", "kotor", "angkat", "wajah", "buat"	“Micellar water pond's baguss banget buat bersihin wajah dari make up seharian.. selain harganya murah kemasannya juga praktis banget buat dibawa kemana-mana dan aku sering pake micellar water ini sehabis aktivitas juga untuk ngangkat debu di wajah aku”	4	4
		“ENAK BANGETT bener2 sensainya lembutt, Ponds Micellar Water oke banget sih buat bersihin muka. Makeup langsung keangkat tanpa perlu gosok keras, plus rasanya ringan di kulit. Gak bikin lengket atau kering, jadi enak dipake sehari-hari. Simple tapi manjur, cocok buat yang suka skincare santai!”	4	4
		“suka banget sama micellar water ponds ini. Dia ini mengandung 5 bahan aktif yang bisa mencerahkan wajah salah satunya niacinamide. Micellar water ini juga ringan banget untuk kulit sensitif aku ini dan cukup ampuh untuk bersihin make up tebel dan yg paling penting dia ini non alcohol luv bngt deh”	4	4
7	"mula", "mata", "eye", "ramah", "double_cleansing", "bawah", "area", "agar", "bersihin", "praktis"	“Enakeun dipake buat bersihin wajah ga ninggalin rasa berminyak diwajah/ malah bikin kering. Bersihnya berasa. Diarea mata ga bikin pedih/ panas. Bikin kulit tetep lembab. Mau repurchase gampang cari dan harganya affordable.”	3	2
		“Please bgt untuk yang suka pakai Garnier Micellar Water, wajib cobain yang rose ini? □□ seenak itu, wangi, ini menurutku lebih bagus daripada yang tutup pink / biru biasanya itu. Memang tidak se oily yang micellar water kuning yg ada oil based nya, tapi pakai ini bener2 bisa ampuh juga untuk ngebersihin area mata yang full eyeshadow bold? ”	3	2

Tabel 4.23 Hasil Validasi Evaluator (Bagian 5)

Topik	Kata Kunci	Sampel Komentar	Evaluator 1	Evaluator 2
		"gampang buat hapus eyeshadow, lem bulu mata, mascara, dan eyebrow pencil. Pakenya juga irit banget gausah banyak banyak, tapi kalo aku kapasnya di tempel ke kelopak mata agak didiemin sambil di press biasa aja pake tangan terus baru dibersihin, hasilnya udh lumayan bersih, tinggal double cleansing pake facial wash."	3	2

Berdasarkan hasil evaluasi pada Tabel 4.23 yang dilakukan terhadap tujuh topik dari model LDA, dapat disimpulkan bahwa sebagian besar topik berhasil merepresentasikan makna secara relevan dan logis sesuai dengan konteks komentar pengguna Female Daily. Untuk memahami persebaran tingkat relevansi komentar berdasarkan hasil evaluasi manual, berikut disajikan distribusi skor dari masing-masing evaluator secara terpisah terhadap 21 komentar.

1. Evaluator 1 memberikan hasil penilaian sebagai berikut:
    a. Skor 2 sebanyak 1 topik,
    b. Skor 3 sebanyak 7 topik,
    c. Skor 4 sebanyak 13 topik.
2. Evaluator 2 memberikan hasil penilaian sebagai berikut:
    a. Skor 2 sebanyak 3 topik,
    b. Skor 3 sebanyak 6 topik,
    c. Skor 4 sebanyak 12 topik.

Jika digabungkan, total skor penilaian yang diberikan oleh kedua evaluator terhadap 21 topik mencapai 42 butir penilaian (2 evaluator × 21 topik). Distribusi skor tersebut menunjukkan bahwa skor 2 diberikan sebanyak 4 kali (9,5%), skor 3 sebanyak 13 kali (31,0%), dan skor 4 sebanyak 25 kali (59,5%). Temuan ini mengindikasikan bahwa mayoritas topik memperoleh tingkat relevansi yang tinggi, sebagaimana terlihat dari tingginya frekuensi skor 4 yang mendominasi hasil evaluasi. Hal ini menunjukkan bahwa topik-topik yang dihasilkan melalui pemodelan topik umumnya dianggap sangat relevan oleh para evaluator dalam merepresentasikan komentar pengguna terhadap produk *skincare*.

Secara keseluruhan, tidak ada topik yang diberi skor 1 (Tidak Relevan), yang menunjukkan bahwa seluruh topik dinilai valid oleh setidaknya satu komentar atau lebih. Ini memperkuat validitas semantik hasil pemodelan topik, menunjukkan bahwa keluaran model

LDA tidak hanya baik secara metrik kuantitatif (misalnya nilai koherensi), tetapi juga kuat secara kualitatif melalui proses interpretasi manusia. Validasi ini selaras dengan pendekatan studi seperti yang diusulkan dalam literatur Puspita dkk. (2024), yang menekankan pentingnya peran *human judgment* dalam mengevaluasi hasil model *unsupervised* agar makna tematik tetap utuh dan dapat digunakan dalam konteks nyata.

Jika dibandingkan dengan studi sebelumnya, seperti penelitian oleh Habibi & Kusumaningtyas (2023) yang menerapkan LDA pada ribuan *tweet* bertagar #*skincare* dan menghasilkan tema seputar bahan alami dan harga, penelitian ini memiliki keunggulan dalam konteks dan kedalaman data. *Dataset* berasal dari lebih dari 66.000 komentar ulasan produk pemenang Female Daily *Best of Beauty Awards*, yang mewakili opini dari komunitas lokal dengan pengalaman riil penggunaan *skincare*. Dengan demikian, topik yang dihasilkan lebih representatif dalam menggambarkan kebutuhan dan persepsi pengguna Indonesia terhadap produk kecantikan.

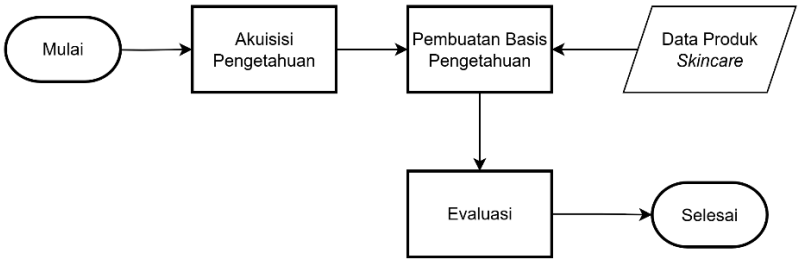
Temuan ini juga sejalan dengan teori *electronic word of mouth* (e-WOM), yang menyoroti bagaimana pengalaman konsumen yang dibagikan secara daring mampu memengaruhi keputusan pembelian orang lain. Topik-topik seperti pengalaman pemakaian, efek, kenyamanan, serta rekomendasi produk mencerminkan praktik e-WOM yang aktif di platform seperti Female Daily. Penelitian yang dilakukan oleh Erdawati dkk. (2023) menyebutkan bahwa e-WOM memainkan peran signifikan dalam membentuk keputusan pembelian *skincare* di kalangan generasi muda, menjadikan validasi semantik seperti ini krusial untuk memastikan model topik benar-benar mencerminkan suara konsumen.

Dengan demikian, validasi ini tidak hanya memberikan penguatan terhadap kualitas topik hasil LDA, tetapi juga membuktikan bahwa pemodelan ini dapat digunakan sebagai representasi konseptual untuk membangun ontologi domain, serta mendukung pemahaman pola komunikasi digital dalam ranah kecantikan.

4.7 Pembuatan Ontologi

Berdasarkan hasil interpretasi *topic modeling* menggunakan metode LDA, diperoleh beragam tema utama yang merepresentasikan pengalaman, kebutuhan, hingga persepsi pengguna terhadap produk *skincare*, mulai dari kandungan bahan aktif, jenis kulit, manfaat produk, hingga masalah kulit yang dihadapi. Topik-topik ini beserta kata kunci dominannya berfungsi sebagai sumber pengetahuan konseptual yang bernilai dalam mendefinisikan struktur domain *skincare* secara lebih sistematis. Oleh karena itu, guna memperluas pemanfaatan hasil

temuan ini, langkah selanjutnya adalah merepresentasikan pengetahuan tematik tersebut ke dalam bentuk ontologi. Pendekatan ontologi dipilih karena memungkinkan transformasi pengetahuan tidak terstruktur menjadi skema konseptual yang formal dan dapat diproses mesin. Dengan demikian, pengembangan ontologi dalam penelitian ini tidak hanya memperkuat kontribusi tematik dari *topic modeling*, tetapi juga membangun kerangka pengetahuan yang terstruktur dan dapat digunakan kembali (*reusable*) untuk berbagai kebutuhan analisis lanjutan di domain *skincare*.



Gambar 4.36 Alur Pembuatan Ontologi

Untuk menggambarkan secara keseluruhan bagaimana ontologi produk *skincare* dibangun dalam penelitian ini, Gambar 4.36 menampilkan alur proses pengembangan ontologi yang dimulai dari tahap akuisisi pengetahuan, yang berfokus pada pengumpulan informasi deskriptif mengenai produk *skincare* pemenang *Best of Beauty Awards* Female Daily. Data yang dikumpulkan mencakup berbagai atribut penting seperti nama produk, *brand*, jenis kulit, manfaat, kandungan utama, dan masalah kulit yang ditangani.

Selanjutnya, pemodelan basis pengetahuan dilakukan untuk mengklasifikasikan elemen-elemen utama ke dalam entitas atau konsep ontologis. Pada tahap ini, ditentukan kelas-kelas seperti Produk, *Brand*, Manfaat, Tipe Kulit, Kandungan Utama, dan Masalah Kulit, yang akan menjadi fondasi struktur ontologi. Informasi yang telah diklasifikasikan kemudian diproses dalam tahap pengolahan data produk *skincare*, yaitu proses representasi entitas ke dalam bentuk formal menggunakan elemen-elemen OWL, seperti *owl:Class*, *owl:ObjectProperty*, dan *owl:DatatypeProperty*. Tahap ini juga melibatkan perumusan domain dan *range* dari masing-masing properti untuk mengatur relasi semantik secara eksplisit. Setelah itu, dilakukan transformasi ke informasi semantik, yaitu pengembangan relasi antar entitas secara eksplisit untuk membentuk struktur pengetahuan konseptual yang saling terhubung. Relasi seperti `diproduksiOleh`, `mengandung`, dan `menanganiMasalah` menggambarkan hubungan logis antar konsep dalam domain *skincare*.

4.7.1 Akuisisi Pengetahuan

Tahap akuisisi pengetahuan bertujuan untuk mengumpulkan dan menstrukturkan informasi penting yang merepresentasikan domain produk *skincare*. Pengetahuan ini diperoleh melalui pengumpulan data deskriptif dan atribut-atribut utama dari produk *skincare* yang terdaftar sebagai pemenang *Best of Beauty Awards* Female Daily dari tahun 2020 hingga 2024. Informasi yang dihimpun mencakup berbagai aspek yang sering dijadikan acuan pengguna dalam memilih *skincare*, seperti jenis produk, *brand*, kandungan utama, jenis kulit yang direkomendasikan, masalah kulit yang ditangani, serta manfaat yang ditawarkan oleh produk.

Pengetahuan-pengetahuan tersebut kemudian diolah menjadi entitas dan konsep yang lebih terstruktur untuk membangun dasar ontologi. Kategori seperti *moisturizer*, *toner*, *sunscreen*, dan *acne treatment* diidentifikasi sebagai tipe produk, sementara nama-nama seperti Somethinc, Avoskin, dan Skintific diklasifikasikan sebagai *brand*. Kandungan aktif seperti *niacinamide*, *salicylic acid*, dan *centella asiatica* juga dimasukkan sebagai entitas penting karena sering muncul dalam formulasi produk dan dikaitkan dengan manfaat spesifik.

Proses akuisisi ini juga mencakup analisis deskriptif terhadap manfaat yang dijanjikan oleh produk, seperti *melembapkan kulit*, *mengurangi jerawat*, *mencerahkan wajah*, dan *mengontrol minyak*. Informasi ini kemudian dikaitkan dengan kategori masalah kulit seperti *kulit kering*, *berjerawat*, *kusam*, atau *sensitif*. Seluruh informasi yang telah diklasifikasikan ini berperan sebagai fondasi awal untuk membangun struktur ontologi yang mencerminkan relasi semantik antar entitas dalam domain *skincare*.

4.7.2 Basis Pengetahuan

Pemodelan basis pengetahuan merupakan tahap penting dalam pembangunan ontologi yang bertujuan untuk membentuk representasi konseptual dari domain produk *skincare* ke dalam struktur formal. Pada penelitian ini, pemodelan dilakukan menggunakan pendekatan *top-down*, di mana proses diawali dengan pendefinisian kelas-kelas utama (*classes*). Setelah itu, dilakukan pengembangan ke tingkat *subclass* dan individu (*instances*) untuk masing-masing entitas yang telah diidentifikasi sebelumnya.

a. Spesifikasi

Sebelum membangun struktur ontologi secara teknis, perlu disusun terlebih dahulu dokumen spesifikasi sebagai acuan dasar pengembangan. Spesifikasi ini berfungsi untuk merumuskan arah dan ruang lingkup ontologi, serta mengidentifikasi kebutuhan informasi

yang hendak dijawab melalui representasi pengetahuan yang dibangun. Penyusunan spesifikasi ini mencakup penjabaran tujuan, cakupan, serta daftar pertanyaan kompetensi yang menjadi dasar evaluasi efektivitas ontologi dalam mendukung pencarian informasi. Rangkuman dari spesifikasi ontologi tersebut disajikan pada Tabel 4.23.

Tabel 4.23 Ontologi Spesifikasi Dokumen

No	Ontologi Spesifikasi Dokumen
1.	**Tujuan**
	Membangun model ontologi untuk memudahkan pengklasifikasian produk *skincare*
2.	**Cakupan**
	Produk *skincare* pemenang *Best of Beauty Awards* 2020 hingga 2024
3.	**Kompetensi Pertanyaan**
	1. Produk apa saja yang mengandung *salicylic acid*? 2. Produk apa saja yang cocok untuk kulit berminyak? 3. Produk apa saja yang menangani jerawat dan cocok untuk kulit berminyak? 4. Produk apa saja yang mengandung *niacinamide* yang membantu mencerahkan kulit untuk masalah kulit kusam?

Dokumen spesifikasi di atas menunjukkan bahwa pengembangan ontologi dilakukan secara terarah dengan memperhatikan kebutuhan pengguna. Pertanyaan kompetensi yang dirumuskan memberikan gambaran konkret mengenai jenis informasi yang ingin diperoleh dari sistem berbasis ontologi. Hal ini sekaligus menjadi landasan dalam menentukan struktur ontologi, mulai dari pemilihan konsep inti (*class*), relasi antar entitas (*property*), hingga individu. Dengan spesifikasi ini, proses pengembangan ontologi dapat dilakukan secara sistematis dan berorientasi pada aplikasi nyata dalam domain produk *skincare.*

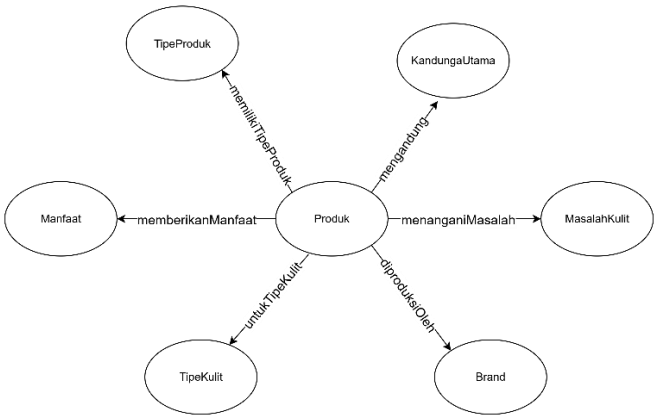
b. Deskripsi

Pembuatan basis pengetahuan pada penelitian ini berfokus pada produk *skincare* pemenang *Best of Beauty Awards* 2020 hingga 2024. Sumber data utama yang digunakan dalam pembangunan ontologi ini adalah informasi deskriptif dari setiap produk pemenang, yang meliputi nama produk, kategori atau tipe produk, merek (*brand*), kandungan utama, jenis kulit yang direkomendasikan, manfaat yang dijanjikan, serta masalah kulit yang ditargetkan. Informasi tersebut diperoleh melalui proses *scraping* dari halaman *review* Female Daily, kemudian diekstraksi dan dibersihkan sehingga membentuk *dataset* terstruktur. *Dataset* inilah yang menjadi basis pengetahuan dalam pembangunan ontologi.

c. Konseptualisasi

Pada tahap ini, struktur konseptual dari ontologi dibangun dalam relasi antar konsep. Berdasarkan hasil *topic modeling*, beberapa kelas utama yang dikembangkan adalah:

1. `Produk`
2. `Brand`
3. `MasalahKulit`
4. `Skintype`
5. `KandunganUtama`
6. `TipeProduk`
7. `Manfaat`



Gambar 4.37 Diagram Konseptual Relasi Antar *Class*

Untuk memperjelas relasi antar kelas dalam ontologi yang telah dirancang, Gambar 4.37 menampilkan diagram konseptual yang menggambarkan hubungan antara entitas utama dalam domain produk *skincare*. Diagram ini merepresentasikan struktur semantik dari ontologi yang dibangun, dengan kelas `Produk` sebagai pusat relasi yang terhubung dengan enam entitas utama lainnya melalui berbagai *object property*.

Tahap konseptualisasi berfokus pada pembangunan struktur konseptual ontologi berdasarkan pengetahuan yang telah diakuisisi. Pada tahap ini, entitas-entitas kunci dalam domain *skincare* diidentifikasi, dikelompokkan ke dalam kelas-kelas, didefinisikan atribut-atributnya, dan ditetapkan hubungan atau relasi antar konsep. Struktur konseptual ini merupakan fondasi logis ontologi dan direpresentasikan menggunakan elemen-elemen dari Web Ontology Language (OWL), yaitu kelas (owl:Class), properti objek (owl:ObjectProperty), properti data (owl:DatatypeProperty), dan individu (owl:NamedIndividual).

1. *Classes*

Classes merepresentasikan konsep-konsep abstrak atau kategori entitas dalam domain `skincare`. Dalam ontologi ini, kelas-kelas utama yang didefinisikan adalah sebagai berikut:

a. `skincare:Produk`: Kelas sentral yang merepresentasikan setiap produk *skincare* individual yang ada dalam ontologi.
b. `skincare:Brand`: Merepresentasikan merek atau perusahaan produsen dari produk *skincare*.
c. `skincare:KandunganUtama`: Merepresentasikan bahan aktif atau komponen kunci yang menjadi formulasi dalam produk *skincare*.
d. `skincare:Manfaat`: Merepresentasikan efek positif atau kegunaan yang dapat diberikan oleh produk *skincare* pada kulit.
e. `skincare:MasalahKulit`: Merepresentasikan berbagai kondisi atau masalah kulit yang dapat ditangani atau diatasi oleh produk *skincare*.
f. `skincare:Skintype`: Merepresentasikan kategori jenis-jenis kulit yang berbeda.
g. `skincare:TipeProduk`: Merepresentasikan kategori atau jenis produk *skincare* berdasarkan bentuk atau fungsinya (misalnya, serum, *moisturizer*, toner).

2. *Object Properties*

Object Properties mendefinisikan jenis hubungan atau relasi semantik antara dua individu (*instances*) dari kelas yang berbeda. Dengan adanya *object properties*, ontologi dapat merepresentasikan hubungan nyata antar konsep dalam dunia produk *skincare*, seperti hubungan antara produk dan manfaat, produk dan bahan aktif, atau produk dengan jenis kulit tertentu.

Tabel 4.24 *Object Property* dan Relasi Antar Konsep (Bagian 1)

Properti Objek	Domain (rdfs:domain)	Range (rdfs:range)	Keterangan
`diproduksiOleh`	`Produk`	`Brand`	Menghubungkan sebuah produk dengan merek yang memproduksinya.
`memberikanManfaat`	`Produk`	`Manfaat`	Menghubungkan sebuah produk dengan manfaat spesifik yang diberikannya.

Tabel 4.24 *Object Property* dan Relasi Antar Konsep (Bagian 2)

Properti Objek	Domain (rdfs:domain)	Range (rdfs:range)	Keterangan
memilikiTipeProduk	Produk	TipeProduk	Menghubungkan sebuah produk dengan kategori atau jenis produknya.
menanganiMasalah	Produk	MasalahKulit	Menghubungkan sebuah produk dengan masalah kulit yang ditanganinya.
mengandung	Produk	KandunganUtama	Menghubungkan sebuah produk dengan bahan aktif yang terkandung di dalamnya.
untukSkintype	Produk	Skintype	Menghubungkan sebuah produk dengan jenis kulit yang direkomendasikan untuknya.

Object properties yang dijelaskan pada Tabel 4.24 ini tidak hanya menjelaskan relasi satu arah, tetapi juga bisa dimanfaatkan dalam *reasoning* semantik. Dengan pemodelan *object property* yang tepat dan lengkap, ontologi ini mampu menyimpan dan mengelola relasi pengetahuan yang kompleks dan saling terkait dalam domain *skincare* secara eksplisit, fleksibel, dan dapat digunakan ulang.

3. *Property Data*

Property data atau *datatype properties* merupakan komponen penting dalam ontologi yang digunakan untuk merepresentasikan atribut literal dari suatu individu (*instance*). Berbeda dengan *object property* yang menghubungkan satu individu dengan individu lain dari kelas berbeda, *property data* menghubungkan individu dengan nilai literal seperti teks (*string*), angka (*integer*, *decimal*), tanggal, atau *boolean*.

Properti ini memungkinkan ontologi menyimpan informasi deskriptif dan data konkret yang dapat digunakan untuk pencarian, pengurutan, ataupun analisis berbasis nilai. Dalam domain *skincare*, data seperti harga produk dan nama produk adalah informasi penting yang tidak direpresentasikan sebagai entitas tersendiri, tetapi sebagai atribut melekat pada suatu individu dari kelas Produk. Dalam pengembangan ontologi ini, digunakan dua *property* data utama yang ditujukan untuk kelas Produk yang bisa dilihat pada Tabel 4.25.

Tabel 4.25 Definisi *Property* Data

Properti Data	Domain (rdfs:domain)	Range (rdfs:range)	Keterangan
`hargaIDR`	`Produk`	xsd:string	Harga produk dalam mata uang Rupiah Indonesia. (Digunakan xsd:string untuk mengakomodasi nilai non-numerik seperti "Tidak tersedia" atau "nan".)
`namaProduk`	`Produk`	xsd:string	Nama lengkap dari produk *skincare*.

4. Individu

Individu adalah entitas konkret atau *instance* dari kelas-kelas yang telah didefinisikan. Pada tahap ini, ontologi dipopulasi dengan data aktual dari produk *skincare* pemenang Female Daily *Best of Beauty Awards* tahun 2020 hingga 2024. Setiap produk, *brand*, bahan aktif, masalah kulit, manfaat, dan jenis kulit direpresentasikan sebagai individu dalam ontologi. Sebagai contoh, beberapa representasi individu dari kelas `skincare:Produk` adalah sebagai berikut:

a. Individu: skincare:AcneAct_Acne_Spot_Gel
   1. Jenis: `skincare:Produk.`
   2. Hubungan dengan Brand: `skincare:diproduksiOleh skincare:Erha.`
   3. Hubungan dengan Manfaat: `skincare:memberikanManfaat skincare:mengeringkan_jerawat, skincare:mengurangi_radang_jerawat, skincare:menyamarkan_bekas_jerawat.`
   4. Hubungan dengan Tipe Produk: `skincare:memilikiTipeProduk skincare:Acne_Treatment.`
   5. Hubungan dengan Masalah Kulit: `skincare:menanganiMasalah skincare:jerawat.`
   6. Hubungan dengan Kandungan Utama: `skincare:mengandung skincare:salicylic_acid, skincare:sulfur.`
   7. Hubungan dengan Jenis Kulit: `skincare:untukSkintype skincare:berjerawat.`

8. Atribut Data: skincare:hargaIDR "65000", skincare:namaProduk "AcneAct Acne Spot Gel".

b. Individu:
skincare:Bright_Miracle_Ultimate_Clarity_Facial_Foam

1. Jenis: skincare:Produk.
2. Hubungan dengan Brand: skincare:diproduksiOleh skincare:Pond_s.
3. Hubungan dengan Manfaat: skincare:memberikanManfaat skincare:memudarkan_bekas_jerawat, skincare:memudarkan_flek, skincare:mencerahkan_kulit, skincare:meratakan_warna_kulit.
4. Hubungan dengan Tipe Produk: skincare:memilikiTipeProduk skincare:Face_Wash.
5. Hubungan dengan Masalah Kulit: skincare:menanganiMasalah skincare:bekas_jerawat, skincare:flek, skincare:kulit_kusam.
6. Hubungan dengan Kandungan Utama: skincare:mengandung skincare:glycerin, skincare:lauric_acid, skincare:niacinamide, skincare:resorcinol, skincare:vitamin_c.
7. Hubungan dengan Jenis Kulit: skincare:untukSkintype skincare:berminyak, skincare:kering, skincare:normal
8. Atribut Data: skincare:hargaIDR "35000", skincare:namaProduk "Bright Miracle Ultimate Clarity Facial Foam"

d. Implementasi

Implementasi ontologi dilakukan menggunakan perangkat lunak Protégé 5.6.5, sebuah *tool* pengembangan ontologi berbasis OWL (*Web Ontology Language*) yang umum digunakan dalam pengembangan sistem berbasis semantik. Pada tahap ini, struktur ontologi yang telah dikonseptualisasikan sebelumnya direalisasikan ke dalam bentuk formal dan dapat diproses mesin. Proses implementasi mencakup pemodelan kelas (*class*), properti (*property*), individu

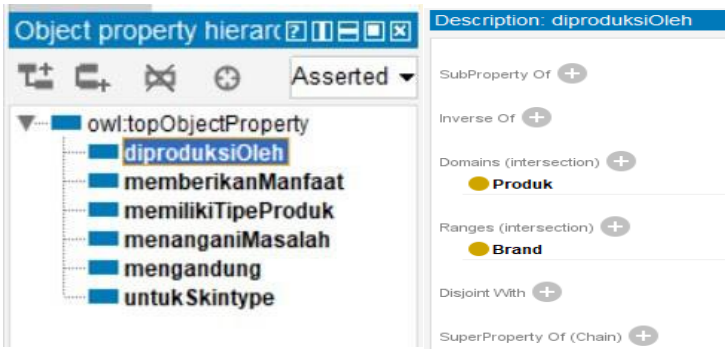
(*instance*), serta verifikasi struktur ontologi menggunakan visualisasi bawaan seperti OntoGraf.

Ontologi ini dimulai dengan kelas induk `owl:Thing`, yang menjadi akar dari seluruh hierarki kelas. Seluruh konsep utama seperti `Brand`, `KandunganUtama`, `Manfaat`, `MasalahKulit`, `Produk`, `Skintype`, dan `TipeProduk` didefinisikan sebagai *subclass* langsung dari `Thing`. Masing-masing kelas ini merepresentasikan entitas utama dalam domain produk *skincare* berdasarkan hasil akuisisi pengetahuan dan konseptualisasi sebelumnya. Struktur kelas divisualisasikan pada Gambar 4.38, yang menunjukkan bagaimana kelas-kelas disusun dalam bentuk hierarki. Penentuan struktur ini penting untuk menjaga keselarasan semantik antar entitas dan sebagai dasar dalam mendefinisikan relasi antar kelas.



Gambar 4.38 Struktur *Class* Ontologi

Setelah menentukan *class* dalam ontologi, langkah selanjutnya menentukan bagaimana *class-class* tersebut saling berhubungan. Hubungan ini direpresentasikan menggunakan *object property*. *Object property* secara eksplisit mendefinisikan relasi antar objek dalam dunia yang dimodelkan. Pembuatan *object property,* domain, dan *range* pada ontologi yang dapat dilihat pada Gambar 4.39.



Gambar 4.39 *Object Property* Ontologi

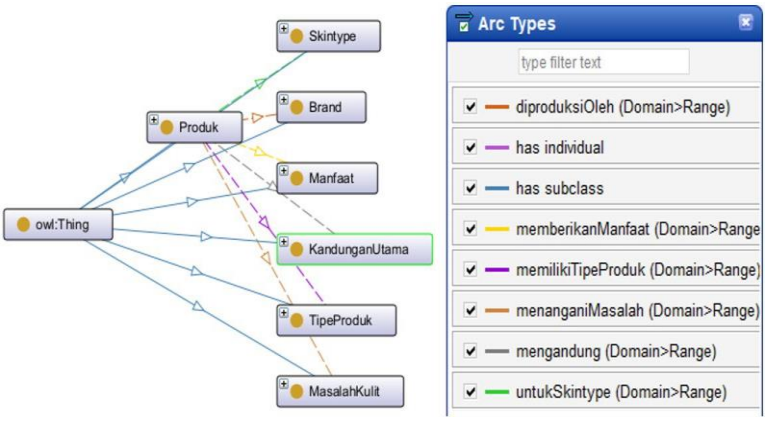
Gambar 4.39 menunjukkan contoh *object property* bernama `diproduksiOleh` dengan domain `Produk` dan *range* `Brand`. Secara sederhana, domain menentukan jenis objek yang memulai relasi (`Produk`), dan *range* menentukan jenis objek yang mengakhiri relasi (`Brand`). Dengan cara ini, ontologi secara otomatis mengerti bahwa hubungan 'diproduksi oleh' hanya dapat terjadi antara sebuah `Produk` dan sebuah `Brand`, menjaga konsistensi dan keakuratan data.

Selanjutnya, pembuatan *instances* yang berhubungan dengan produk *skincare.* Pada Gambar 4.40 menampilkan *instances* yang dibangun untuk *class* `Brand`. *Instance* ini merepresentasikan contoh merek produk *skincare* pemenang Female Beauty *Best of Beauty Awards* periode 2020 hingga 2024. Secara total, telah dibuat 2.417 *instances* yang mencakup berbagai produk, *brand*, bahan aktif, dan lain-lain.



Gambar 4.40 *Instances* Ontologi

Untuk memastikan struktur ontologi telah dibangun dengan baik dan relasi antar elemen konsisten, digunakan fitur visualisasi OntoGraf yang tersedia di dalam Protégé. Gambar 4.41 menunjukkan hasil visualisasi hubungan antar kelas, properti, dan individu. Dengan OntoGraf, pengguna dapat melihat bagaimana entitas saling terhubung, memvalidasi struktur relasi, serta menelusuri jalur relasional dari satu individu ke individu lainnya. Visualisasi ini memudahkan proses verifikasi internal untuk mengecek kelengkapan, konsistensi, dan kebermaknaan semantik dari ontologi yang dibangun, sekaligus mendukung dokumentasi teknis dalam laporan penelitian.



Gambar 4.41 Visualisasi dengan Ontograf

Ontologi disimpan dalam format OWL (.owl) menggunakan sintaks RDF/XML, sehingga dapat digunakan lebih lanjut pada sistem berbasis semantik, *reasoning engine*, atau aplikasi berbasis web semantik seperti SPARQL *endpoint*. Format ini memastikan ontologi dapat diintegrasikan dengan *tool* analisis lanjutan atau digunakan kembali dalam pengembangan sistem cerdas pada domain *skincare*. Data dasar untuk penyusunan ontologi disajikan pada Lampiran D.

4.7.3 Evaluasi

Tahap evaluasi bertujuan untuk memastikan bahwa ontologi yang dibangun telah sesuai dengan kebutuhan, baik dari segi struktur, relasi antar konsep, maupun kelengkapan pengetahuan yang diwakili. Evaluasi dilakukan dengan pengujian fungsional berbasis *query*. Dalam tahap ini, serangkaian pertanyaan dirumuskan, yang relevan dengan domain aplikasi dan kebutuhan pengguna, khususnya terkait data produk *skincare* yang menjadi pemenang penghargaan Female Daily BoBA. Pertanyaan-pertanyaan ini berfungsi sebagai skenario pengujian untuk melihat sejauh mana ontologi mampu menyimpan dan menyediakan informasi yang dibutuhkan. Jawaban atas pertanyaan-pertanyaan tersebut kemudian diperoleh melalui eksekusi *query* menggunakan SPARQL. Berikut disajikan daftar pertanyaan yang berfungsi sebagai skenario untuk menguji kapabilitas ontologi *skincare*.

Tabel 4.26 Pertanyaan Pengujian Ontologi *Skincare*

No	Pertanyaan
1	Produk apa saja yang mengandung *salicylic acid*?
2	Produk apa saja yang cocok untuk kulit berminyak?
3	Produk apa saja yang menangani jerawat dan cocok untuk kulit berminyak?
4	Produk apa saja yang mengandung *niacinamide* yang membantu mencerahkan kulit untuk masalah kulit kusam?

Berdasarkan daftar pertanyaan pada Tabel 4.26, maka akan dilakukan pengujian ontologi menggunakan SPARQL *query*. Setiap *query* dikembangkan dengan tujuan untuk mengekstraksi informasi produk *skincare* yang relevan dari ontologi yang telah dimodelkan. Struktur dasar setiap *query* meliputi identifikasi entitas sebagai "Produk" (`produkUri rdf:type skincare:Produk`) dan pengambilan atribut nama produk (`skincare:namaProduk ?namaProduk`). Selanjutnya, kriteria spesifik untuk setiap pertanyaan penelitian akan diimplementasikan sebagai kondisi tambahan untuk memfilter hasil yang paling relevan.

1. Pertanyaan 1: Produk apa saja yang mengandung *salicylic acid*?

SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX skincare: <http://skincare.example.org/ontology/>

SELECT (STR(?namaProduk) AS ?NamaProduk)
WHERE {
 ?produkUri rdf:type skincare:Produk ;
        skincare:namaProduk ?namaProduk ;
        skincare:mengandung skincare:salicylic_acid .
}
```

NamaProduk
"Sébium Serum Anti-Acne & Anti-Aging Dermatological Serum"
"Real Mugwort Salicylic Acid Acne Control Clay Mask Stick"
"Water Rose Micellar Cleansing Water"
" Acneplus Spot Care Pro Serum"
"AcneAct Acne Spot Gel"
"Acne Warrior Paste"
"Glow Maker AHA BHA PHA Clarifying Treatment Toner"
"Mugwort Acne Clay Stick"
" Revitalift Crystal Gel Cream"
" Hyamuno Contr-Oil Oil Control Facial Wash"
"Acne Cream"
"Micellar Salicylic BHA"
"Mugwort Pore Clarifying Mask"

Gambar 4.42 Hasil Query Pertanyaan 1

Pada Gambar 4.42 menyajikan implementasi SPARQL *query* yang dirancang untuk mengekstraksi informasi nama produk *skincare* yang mengandung *salicylic acid*. Untuk mencapai tujuan ini, kondisi `skincare:mengandung skincare:salicylic_acid` ditambahkan pada *query*. Kondisi ini berfungsi sebagai filter utama yang memastikan bahwa hanya produk-produk yang secara spesifik memiliki relasi `mengandung` dengan `salicylic_acid` dalam ontologi yang ditampilkan sebagai hasil.

2. Pertanyaan 2:

SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX skincare: <http://skincare.example.org/ontology/>

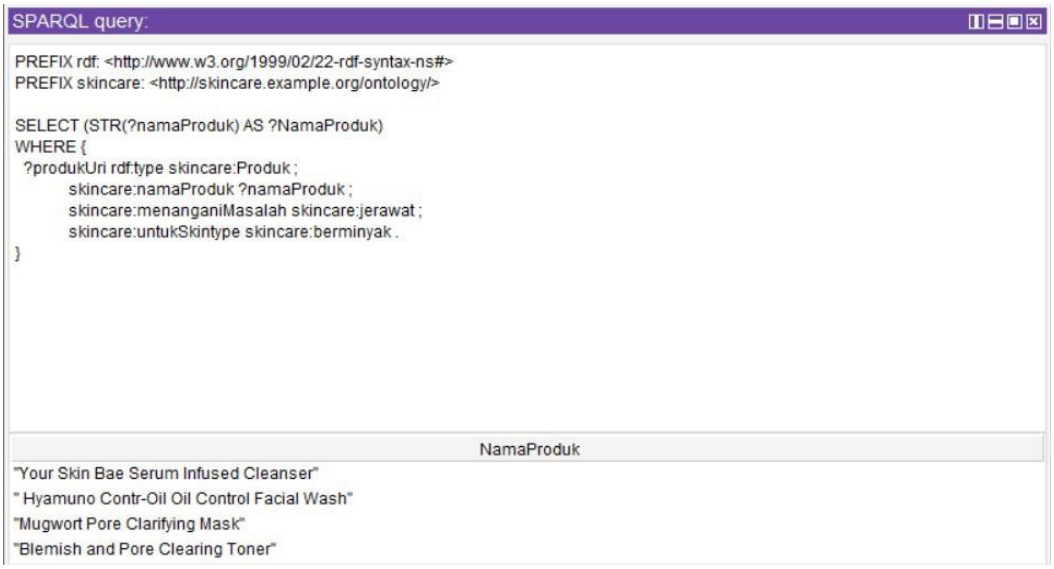
SELECT (STR(?namaProduk) AS ?NamaProduk)
WHERE {
 ?produkUri rdf:type skincare:Produk ;
        skincare:namaProduk ?namaProduk ;
        skincare:untukSkintype skincare:berminyak .
}
```

NamaProduk
"Radiant Skin Serum"
" Hyamuno Contr-Oil Oil Control Facial Wash"
"Glycolic Bright Peeling Toner"
"Refreshing Face Mist Antibacterial"
"Hydra Rose Gel-to-Foam Cleanser"
"Radian-C Double Active Brightening Essence"
"Lip Therapy Stick"
"Evolution of Goodness Face Oil"
"Revitalift Pro Youth Face Mask"
"Skin Barrier Water Cream"
" Revitalift Crystal Gel Cream"
"Your Skin Bae Serum Lactic Acid 10% + Kiwi Fruit 5% + Niacinamide 2.5%"
"Gentle Clean! Deep Cleansing Balm-Cocomelt"
"Your Skin Bae Serum Alpha Arbutin 3% + Grapeseed"

Gambar 4.43 Hasil Query Pertanyaan 2

Pada Gambar 4.43 menyajikan implementasi SPARQL *query* yang dirancang untuk mengekstraksi nama produk yang cocok untuk kulit berminyak. Untuk mencapai tujuan ini, kondisi `skincare: untukSkintype skincare:berminyak` ditambahkan pada *query*. Kondisi ini berfungsi sebagai filter utama yang memastikan bahwa hanya produk-produk yang secara spesifik memiliki hubungan `untukSkintype` dengan `berminyak` dalam ontologi yang ditampilkan sebagai hasil.

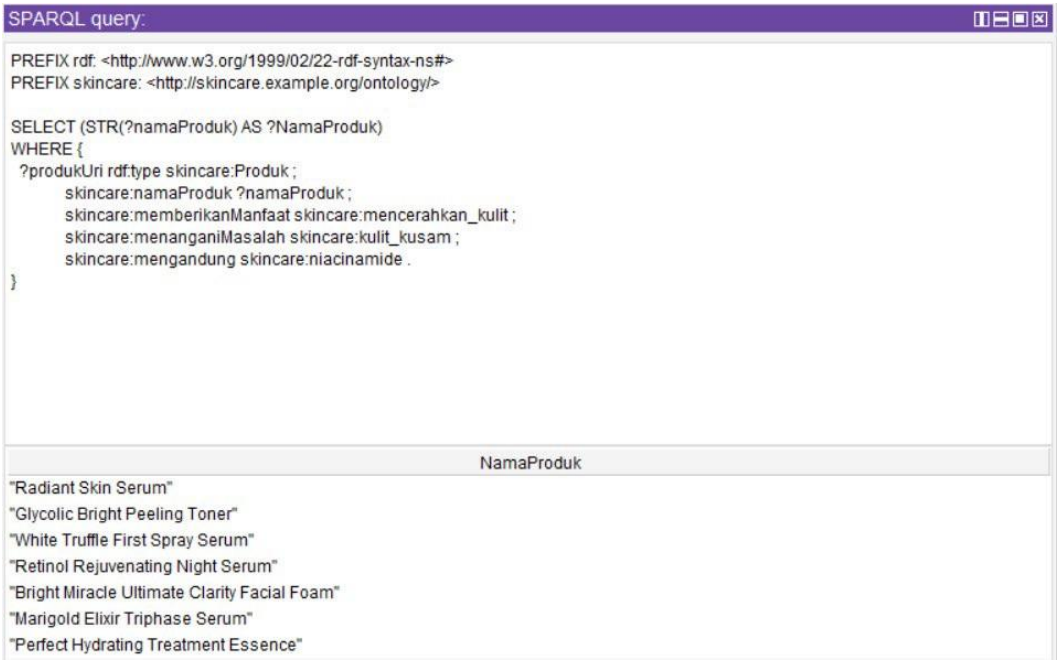
3. Pertanyaan 3:



Gambar 4.44 Hasil Query Pertanyaan 3

Pada Gambar 4.44 menyajikan implementasi SPARQL *query* yang dirancang untuk mengekstraksi nama produk yang menangani jerawat dan cocok untuk kulit berminyak. Query ini menggunakan dua kondisi filter. Kondisi pertama, `skincare:menanganiMasalah skincare:jerawat;` untuk memastikan bahwa hanya produk-produk yang secara spesifik memiliki hubungan `menanganiMasalah` dengan `jerawat`. Kondisi kedua, `skincare:untukSkintype skincare:berminyak` untuk memastikan bahwa hanya produk-produk yang secara spesifik memiliki hubungan `untukSkintype` dengan `berminyak`.

4. Pertanyaan 4:



Gambar 4.45 Hasil Query Pertanyaan 4

Pada Gambar 4.45 menyajikan implementasi SPARQL *query* yang dirancang untuk mengekstraksi nama produk yang menangani jerawat dan cocok untuk kulit berminyak. Query ini menggunakan tiga kondisi filter. Kondisi pertama, `skincare:memberikanManfaat skincare:mencerahkan_kulit` untuk memfilter produk-produk yang `memberikanManfaat` untuk `memberikanManfaat`. Kondisi kedua, `skincare:menanganiMasalah skincare:kulit_kusam;` untuk memastikan bahwa hanya produk-produk yang secara spesifik memiliki hubungan `menanganiMasalah` dengan `kulit_kusam`. Kondisi terakhir, `skincare: mengandung skincare:niacinamide` untuk membatasi hasil pada produk-produk yang `mengandung` kandungan utama yaitu `niacinamide`.

Pembuatan ontologi dalam domain *skincare* pada penelitian ini mampu menghasilkan representasi formal dari pengetahuan yang diperoleh dari hasil *topic modeling* ulasan pengguna di platform Female Daily. Ontologi yang dibangun mencakup entitas penting seperti `Produk`, `Brand`, `TipeProduk`, `Skintype`, `MasalahKulit`, `KandunganUtama`, dan `Manfaat`, serta relasi-relasi semantik di antaranya. Dengan menggunakan pendekatan berbasis proses dan memanfaatkan editor ontologi Protégé, struktur ontologi berhasil

dikembangkan dalam format OWL yang dapat digunakan untuk mendukung pemrosesan semantik lanjutan, seperti sistem rekomendasi atau pencarian berbasis makna. Hasil ini menunjukkan bahwa ontologi dapat berperan sebagai jembatan antara analisis data berbasis teks dengan pengembangan sistem pengetahuan yang dapat dipahami oleh mesin.

Penelitian ini memiliki beberapa kelebihan yang dapat dijadikan dasar pengembangan lebih lanjut. Salah satunya adalah penggunaan data ulasan asli dari platform Female Daily yang kaya konteks dan mencerminkan opini konsumen secara organik. Pendekatan *topic modeling* dengan metode LDA juga terbukti mampu mengidentifikasi topik-topik semantik yang relevan dan dapat diinterpretasikan dengan baik oleh evaluator. Proses ini dilengkapi dengan tahap validasi berbasis *human judgment* untuk memastikan kualitas hasil topik. Lebih lanjut, pengembangan ontologi dari hasil topik memberikan nilai tambah berupa struktur pengetahuan eksplisit yang dapat dimanfaatkan dalam berbagai aplikasi berbasis semantik.

Namun demikian, terdapat beberapa kekurangan yang perlu menjadi bahan evaluasi. Salah satunya adalah proses *preprocessing* yang dilakukan secara semi-manual, yang dapat menimbulkan bias atau inkonsistensi dalam pembersihan data. Selain itu, pemilihan jumlah topik dalam LDA masih bersifat eksperimental tanpa pendekatan otomatis untuk optimasi jumlah topik terbaik. Keterbatasan lainnya adalah cakupan data yang masih terbatas pada satu platform saja. Sebagai tindak lanjut, penelitian ke depan disarankan untuk mengembangkan pipeline *preprocessing* yang lebih otomatis dan konsisten, menggunakan pendekatan *topic coherence optimization*, serta memperluas sumber data dari berbagai platform seperti Shopee, TikTok, atau media sosial lainnya. Ontologi yang dibangun juga dapat diintegrasikan ke dalam sistem rekomendasi atau aplikasi berbasis semantik guna meningkatkan manfaat praktisnya.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Penelitian ini berhasil memetakan topik-topik utama dalam komentar pengguna terhadap produk *skincare* pemenang penghargaan Female Daily *Best of Beauty Awards* (BoBA) dengan menggunakan metode *Latent Dirichlet Allocation* (LDA). Proses ini melibatkan tahapan pengumpulan data, *preprocessing*, dan *topic modeling*, diperoleh sejumlah topik yang mencerminkan perhatian dan pengalaman pengguna terhadap berbagai aspek produk, seperti efektivitas, kecocokan dengan jenis kulit, harga, kemasan, dan aroma. Dengan demikian, penerapan LDA dalam penelitian ini terbukti efektif untuk mengidentifikasi tema-tema utama dalam ulasan pengguna secara otomatis dan sistematis. Model LDA terbaik ditentukan lewat nilai *coherence*, di mana model dengan tujuh topik menunjukkan hasil paling optimal sebesar 0,447733. Selain itu, visualisasi menggunakan `pyLDAvis` berhasil menyajikan representasi visual dari distribusi topik yang dominan dengan memperlihatkan sebaran kata-kata kunci yang saling berkaitan dalam setiap topik mengenai produk *skincare*.

Selain itu, temuan dalam ulasan pengguna juga menunjukkan adanya praktik *electronic word of mouth* (e-WOM) yang kuat, di mana konsumen saling berbagi pengalaman, rekomendasi, dan evaluasi produk secara terbuka melalui platform Female Daily. Informasi yang terkandung dalam e-WOM ini tidak hanya mencerminkan persepsi dan kebutuhan konsumen, tetapi juga berperan penting dalam memengaruhi keputusan pembelian calon pengguna lainnya. Dengan demikian, analisis e-WOM melalui pemodelan topik memberikan wawasan yang lebih dalam tentang pola komunikasi konsumen dan potensi pengaruh sosialnya dalam ranah digital.

Berdasarkan hasil pemodelan topik tersebut dan kebutuhan akan struktur pengetahuan yang lebih terorganisir, ontologi produk *skincare* berhasil dibangun untuk memudahkan pengklasifikasian dan ekstraksi informasi yang relevan. Proses ini diawali dengan perancangan ontologi yang meliputi pembuatan kelas dan *subclass*, serta pendefinisian *data property*, domain, dan *range*. Spesifikasi ontologi difokuskan pada produk *skincare* pemenang Female Daily Beauty Awards 2019-2024 dengan tujuan membangun model ontologi mandiri. Tahap akuisisi pengetahuan dilakukan dengan mengolah data komentar produk *skincare* dari situs Female Daily melalui *topic modeling* menggunakan model LDA. Hasil *topic modeling* ini kemudian diinterpretasikan untuk mengidentifikasi entitas domain dan menghasilkan daftar

kata kunci representatif yang menjadi dasar konseptualisasi struktur ontologi. Kelas-kelas utama seperti `Produk`, `Brand`, `TipeProduk`, `Skintype`, `MasalahKulit`, `KandunganUtama`, dan `Manfaat` dikembangkan, masing-masing dengan *subclass*, atribut, dan relasinya.

Implementasi ontologi dilakukan menggunakan Protege 5.6.5, yang melibatkan definisi hierarki kelas, *object property,* dan instansiasi data merek. Visualisasi dengan Ontograf juga memperjelas relasi ini. Evaluasi ontologi dilakukan melalui pengujian fungsional berbasis *query* SPARQL. Serangkaian pertanyaan relevan diajukan sebagai skenario pengujian. Setiap *query* dikembangkan untuk mengekstraksi informasi spesifik. Hasil pengujian menunjukkan bahwa ontologi yang dibangun mampu menyimpan dan menyediakan informasi yang dibutuhkan secara konsisten dan akurat sesuai domain aplikasi. Secara keseluruhan, penelitian ini memberikan kontribusi dalam memahami persepsi konsumen dari ulasan *online* dan menjadi mampu menyajikan pengetahuan secara terstruktur dan semantik, sehingga mempermudah pengguna dalam menemukan produk *skincare* yang sesuai dengan kebutuhannya pada domain *skincare* di platform Female Daily.

5.2 Saran

Berdasarkan analisis mendalam yang dilakukan terhadap hasil dan pembahasan yang diperoleh dari penelitian ini, penelitian ini menunjukkan bahwa jumlah data yang dianalisis mencapai 66.521 komentar menjadikan pemrosesan cukup kompleks dan memakan waktu. Oleh karena itu, disarankan agar penelitian selanjutnya memanfaatkan pendekatan otomatisasi yang lebih optimal dan efisien untuk tahap pembersihan data.

DAFTAR PUSTAKA

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/frai.2020.00042

Amin, M., Rezaei, S., & Abolghasemi, M. (2014). User satisfaction with mobile websites: the impact of perceived usefulness (PU), perceived ease of use (PEOU) and trust. *Nankai Business Review International*, *5*(3), 258–274. https://doi.org/10.1108/NBRI-01-2014-0005

Awangga, R. M., Prastya, R., Simamora, T. A. D., Majesty, D., & Fadilah, I. (2019). *Dasar-dasar Python*. Kreatif. https://books.google.co.id/books?id=YpzDDwAAQBAJ

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Cahyono, N., Utami, E., & Amborowati, A. (2016). Konseptualisasi Ontologi Iklan Website. *Jurnal Buana Informatika*, *7*(4). https://doi.org/https://doi.org/10.24002/jbi.v7i4.768

Chatterjee, A., Prinz, A., Gerdes, M., & Martinez, S. (2021). An Automatic Ontology-Based Approach to Support Logical Representation of Observable and Measurable Data for Healthy Lifestyle Management: Proof-of-Concept Study. *Journal of Medical Internet Research*, *23*(4), e24656. https://doi.org/10.2196/24656

Chilmi, M. L. C. (2021). *Latent Dirichlet Allocation (LDA) Untuk Mengetahui Topik Pembicaraan Warganet Twitter Tentang Omnibus Law* [Univeristas Islam Negeri Syarif Hidayatullah Jakarta]. https://repository.uinjkt.ac.id/dspace/handle/123456789/56724

Destarani, A. R., Slamet, I., & Subanti, S. (2019). Trend Topic Analysis using Latent Dirichlet Allocation (LDA) (Study Case: Denpasar People's Complaints Online Website). *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika*, *5*(1). https://doi.org/10.26555/jiteki.v5i1.13088

Elwalda, A., Lü, K., & Ali, M. (2016). Perceived Derived Attributes of Online Customer Reviews. *Computers in Human Behavior*, *56*, 306–319. https://doi.org/10.1016/j.chb.2015.11.051

Erdawati, Endarwita, & Widiasari, R. (2023). THE INFLUENCE OF ELECTRONIC WORD OF MOUTH (e-WOM) AND PRODUCT QUALITY ON SKINCARE PURCHASE

DECISIONS (Case Study of Generation Y in Lubuk Sikaping). *Journal of Social and Economics Research*, *5*(1), 184–190. https://idm.or.id/JSER/index.

Fahrezi, I. A., Rudiman, & Verdikha, N. A. (2024). Analisis Sentimen Twitter Atas Isu Hak Angket Menggunakan Pembobotan TF-IDF dan Algoritma SVM. *Sci-Tech Journal*, *3*, 179–192. https://doi.org/10. 56709/stj.v3i2.526

Female Daily. (t.t.). *Female Daily: About Us*. Diambil 12 Mei 2024, dari https://femaledaily.com/about

Figuera, P., & Bringas, G. P. (2024). Revisiting Probabilistic Latent Semantic Analysis: Extensions, Challenges and Insights. Dalam *Technologies* (Vol. 12, Nomor 1). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/technologies12010005

Habibi, M., & Kusumaningtyas, K. (2023). Customer Experience Analysis Skincare Products Through Social Media Data Using Topic Modeling and Sentiment Analysis. *JOURNAL OF SCIENCE AND APPLIED ENGINEERING*, *6*(1), 1. https://doi.org/10.31328/jsae.v6i1.4169

Harits, M. (2016). *Kisah Sukses Hanifa Ambadar dan Affi Asegaf Membawa Female Daily Menjadi Forum Wanita Terbesar di Indonesia*. https://ilovelife.co.id/blog/kisah-sukses-hanifa-ambadar-dan-affi-asegaf-membawa-female-daily-menjadi-forum-wanita-terbesar-di-indonesia/.

Himawan, Harjanti, T. W., Supriati, R., & Setiyani, H. (2020). Evolusi Penggunaan Teknologi Web 3.0 : Semantic Web. *Journal of Information System,Graphics, Hospitality and Technology*, *2*(2), 54–60. https://doi.org/https://doi.org/10.37823/insight.v2i02.107

Jayadianti, H., Damayanti, R., & Juwairiah. (2020). Latent Semantic Analysis (LSA) Dan Automatic Text Summarization (ATS) Dalam Optimasi Pencarian Artikel COVID 19. *Seminar Nasional Informatika 2020 (SEMNASIF 2020)*, *1*(1).

Kementerian Koordinator Bidang Perekonomian Republik Indonesia. (2024, Februari 3). *Hasilkan Produk Berdaya Saing Global, Industri Kosmetik Nasional Mampu Tembus Pasar Ekspor dan Turut Mendukung Penguatan Blue Economy*. https://www.ekon.go.id/publikasi/detail/5626/hasilkan-produk-berdaya-saing-global-industri-kosmetik-nasional-mampu-tembus-pasar-ekspor-dan-turut-mendukung-penguatan-blue-economy

Kherwa, P., & Bansal, P. (2020). Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems*, *7*(24), 1–16. https://doi.org/10.4108/eai.13-7-2018.159623

Khotimmah, N., Sri, D., Giningroem, W. P., Pitoyo, B. S., Ekonomi, F., Bisnis, D., Bhayangkara, U., & Raya, J. (2024). Pengaruh Citra Merek, Kualitas Produk, Online Customer Review, Dan Online Customer Rating Terhadap Keputusan Pembelian Produk Skincare Madame Gie (Studi Kasus Pada Mahasiswa Prodi Manajemen Universitas Bhayangkara Jakarta Raya Angkatan 2020). Dalam *Indonesian Journal of Economics and Strategic Management (IJESM* (Vol. 2, Nomor 1).

Kurniawan, T. A., Safa'at, R., Santoso, N., & Ismail, M. F. (2024). An ontology-based knowledge representation using OWL for Indonesian local regulations. *Indonesian Journal of Electrical Engineering and Computer Science*, *34*(2), 983–990. https://doi.org/10.11591/ijeecs.v34.i2.pp983-990

Mohammed, S. H., & Al-Augby, S. (2020). LSA & LDA topic modeling classification: Comparison study on E-books. *Indonesian Journal of Electrical Engineering and Computer Science*, *19*(1), 353–362. https://doi.org/10.11591/ijeecs.v19.i1.pp353-362

Muryanti, N. L., Virginia, G., Susanto, B., & Proboyekti, U. (2021). Pembangunan Model Pengetahuan Kerajinan Tradisional Indonesia dengan Pendekatan On-To-Knowledge. *Jurnal Terapan Teknologi Informasi*, *4*, 65–75. https://doi.org/https://doi.org/10.21460/jutei.2020.42.196

Mustahidah, M. (2021). *Topic Modeling pada Ulasan Hotel Menggunakan Latent Dirichlet Allocation (LDA) dan Probabilistic Latent Semantic Analysis (PLSA)*. http://repository.its.ac.id/id/eprint/91644

Nanyonga, A., & Wild, G. (2025). Analyzing Aviation Safety Narratives with LDA, NMF and PLSA: A Case Study Using Socrata Datasets. *2025 2nd International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)*, 1–6. https://doi.org/10.1109/RMKMATE64874.2025.11042356

Novarian, N., Khomsah, S., & Arifa, A. B. (2023). Topic Modeling Tugas Akhir Mahasiswa Fakultas Informatika Institut Teknologi Telkom Purwokerto Menggunakan Metode Latent Dirichlet Allocation. *LEDGER: Journal Informatic and Information Technology*, *2*(1), 14–27.

Nugumanova, A., Akhmed-Zaki, D., Mansurova, M., Baiburin, Y., & Maulit, A. (2022). NMF-based approach to automatic term extraction. *Expert Systems with Applications*, *199*, 117179. https://doi.org/10.1016/j.eswa.2022.117179

Patricia, G. (2021). *Ekstraksi Topik Review Tamu Hotel Menggunakan Latent Dirichlet Allocation dan Non-Negative Matrix Factorization* [Institut Teknologi Sepuluh Nopember]. http://repository.its.ac.id/id/eprint/91305

Peraturan Menteri Kesehatan. (2010). Peraturan Menteri Kesehatan Nomor 1175/MENKES/PER/VIII/2010 Tahun 2010 tentang Izin Produksi Kosmetika. Dalam *https://regulasi.bkpk.kemkes.go.id/detail/f6116194-4a7c-4696-a8a4-fbdcf3cb5507/*. https://peraturan.bpk.go.id/Details/129878/permenkes-no-1175menkesperviii2010-tahun-2010

Pramartha, C. (2020). Pengembangan Ontologi Tujuan Wisata Bali dengan Pendekatan Kulkul Knowledge Framework. *SINTECH (Science and Information Technology) Journal*, *3*, 77–89. https://doi.org/https://doi.org/10.31598/sintechjournal.v3i2.592

Puspita, E., Shiddieq, D. F., & Roji, F. F. (2024). Pemodelan Topik pada Media Berita Online Menggunakan Latent Dirichlet Allocation (Studi Kasus Merek Somethinc). *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(2), 481–489. https://doi.org/10.57152/malcom.v4i2.1204

R Wahyudi, M. D., Fatwanto, A., Kiftiyani, U., & Galih Wonoseto, M. (2021). Topic Modeling of Online Media News Titles during COVID-19 Emergency Response in Indonesia Using the Latent Dirichlet Allocation (LDA) Algorithm. *Telematika*, *14*(2), 101–111. https://doi.org/10.35671/telematika.v14i2.1225

Rachman, D. D., & Nurjanah, D. (2019). Sistem Rekomendasi Berbasis Model Ontology dalam Memberikan Rekomendasi Diet Makanan untuk Penderita Penyakit Diabetes. *eProceedings of Engineering*, *6*(2).

Rahman, S., Sembiring, A., Siregar, D., Khair, H., Prahmana, G., Puspadini, R., & Zen, M. (2023). *PYTHON : DASAR DAN PEMROGRAMAN BERORIENTASI OBJEK TAHTA MEDIA GROUP*. Penerbit Tahta Media. https://tahtamedia.co.id/index.php/issj/article/view/344

Rosalinda, G., Santoso, R., & Kartikasari, P. (2023). PEMODELAN TOPIK ULASAN APLIKASI NETFLIX PADA GOOGLE PLAY STORE MENGGUNAKAN LATENT DIRICHLET ALLOCATION. *Jurnal Gaussian*, *11*(4), 554–561. https://doi.org/10.14710/j.gauss.11.4.554-561

Santosa, I., Wiwaha, P. Y., & Suteja, B. R. (2023). Penerapan Model Semantik dalam Pengolahan Arsip Berita. *JUSTINDO (Jurnal Sistem dan Teknologi Informasi Indonesia)*, *8*(2), 75–86. https://doi.org/10.32528/justindo.v8i2.230

Setijohatmo, U. T., Rachmat, S., Susilawati, T., Rahman, Y., Komputer, J. T., Bandung, N., Gegerkalong Hilir, J., & Ciwaruga, D. (2020). Analisis Metoda Latent Dirichlet Allocation untuk Klasifikasi Dokumen Laporan Tugas Akhir Berdasarkan Pemodelan Topik. *Prosiding Industrial Research Workshop and National Seminar*, *11*(1), 402–408. https://doi.org/https://doi.org/10.35313/irwns.v11i1.2040

Sualang, G. M. C., Mandey, S. L., & Ogi, I. W. J. (2023). PENGARUH ELECTRONIC WORD OF MOUTH (E-WOM), HARGA, DAN KUALITAS PRODUK TERHADAP KEPUTUSAN PEMBELIAN PADA PT. MANDALA FINANCE Tbk, CABANG RATAHAN. *JURNAL EMBA: JURNAL RISET EKONOMI, MANAJEMEN, BISNIS DAN AKUNTANSI*, *11*(2), 335–344.

Supono, R. A., & Suprayogi, M. S. (2021). Perbandingan Metode TF-ABS dan TF-IDF Pada Klasifikasi Teks Helpdesk Menggunakan K-Nearest Neighbor. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *5*(5), 911–918. https://doi.org/10.29207/resti.v5i5.3403

Thoumrungroje, A. (2014). The Influence of Social Media Intensity and EWOM on Conspicuous Consumption. *Procedia - Social and Behavioral Sciences*, *148*, 7–15. https://doi.org/10.1016/j.sbspro.2014.07.009

Tondang, B. A., Muhammad Rizqan Fadhil, Muhammad Nugraha Perdana, Akhmad Fauzi, & Ugra Syahda Janitra. (2023). Analisis pemodelan topik ulasan aplikasi BNI, BCA, dan BRI menggunakan latent dirichlet allocation. *INFOTECH : Jurnal Informatika & Teknologi*, *4*(1), 114–127. https://doi.org/10.37373/infotech.v4i1.601

Wahyu, R., & Putri, D. W. (2023). Pengaruh Electronic Word of Mouth Website Female Daily terhadap Keputusan Followers Membeli Produk Kosmetik. *Bandung Conference Series: Public Relations*, *3*(2), 941–948. https://doi.org/10.29313/bcspr.v3i2.9417

Waluyo, D. (2023, Desember 3). *Fenomena Cantik Industri Kosmetik*. https://indonesia.go.id/kategori/editorial/7804/fenomena-cantik-industri-kosmetik?lang=1

Wu, P., An, M., Zou, H. R., Zhong, C. Y., Wang, W., & Wu, C. P. (2020). A robust semi-supervised NMF model for single cell RNA-seq data. *PeerJ*, *8*. https://doi.org/10.7717/peerj.10091

Yutika, C. H., Adiwijaya, A., & Faraby, S. Al. (2021). Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(2), 422. https://doi.org/10.30865/mib.v5i2.2845

Zoya, Latif, S., Shafait, F., & Latif, R. (2021). Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling. *IEEE Access*, *9*, 127531–127547. https://doi.org/10.1109/ACCESS.2021.3112620

LAMPIRAN

Lampiran A - Data Hasil *Scraping*
https://docs.google.com/spreadsheets/d/1oZj8cMuY5jpvI1-f9B4RPfxOyx1Ic_bj/edit?usp=drive_link&ouid=114993123721138206047&rtpof=true&sd=true

Lampiran B - Kamus *Normalization*
https://docs.google.com/spreadsheets/d/1AnrhXfp5GTWB1XXyfReQAr2nImcp44x3/edit?usp=drive_link&ouid=114993123721138206047&rtpof=true&sd=true

Lampiran C - Daftar Stopword
https://drive.google.com/file/d/1eiVXjabz7JI_YWWfaOW9uDPZ7M0UOTSR/view?usp=drive_link

Lampiran D – Data untuk Penyusunan Ontologi
https://docs.google.com/spreadsheets/d/1Gi1fLXSRrJXEOdcdW0N1Uom8P0RoJ0rJ/edit?usp=drive_link&ouid=114993123721138206047&rtpof=true&sd=true