PENGEMBANGAN MODEL ANALISA DISKUSI TOPIC PRODUK MENGGUNAKAN LATENT DIRICHLET ALLOCATION PADA PLATFORM YOUTUBE

Farrikh Alzami*, **Dwi Puji Prabowo, Puri Sulistiyawati, Ahmad Akrom, Rama Aria Megantara, Ricardus Anggi Pramunendar, Dewi Pergiwati**
Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang

*Email korespodensi: alzami@dsn.dinus.ac.id

Abstract.
The increasing number of people engaging in shopping transactions is driven by the rapid development of technology and the ongoing COVID-19 pandemic. When purchasing products, prospective customers typically seek information from various sources, including YouTube. Content creators often review products, and viewers provide comments regarding those products. In this study, researchers will employ the Latent Dirichlet Allocation (LDA) method to expedite the information absorption process regarding these products. Consequently, prospective customers can more easily identify products to purchase based on the reviews from content creators and viewers, categorized by identified topics through clustering. The results of the LDA can also be utilized by the brand owners to assess the level of acceptance of their products among the public. The final outcome of this research is the design of an LDA model for customer topics.

Keywords: topic clustering; review; Latent Dirichlet Allocation; Youtube

Abstrak

Meningkatnya jumlah masyarakat yang melakukan transaksi belanja didasari oleh kemajuan teknologi yang berkembang pesat dan adanya pandemi covid-19 ini. Untuk membeli produk, biasanya calon konsumen akan mencari informasi dari berbagai sumber, tidak terkecuali dari Youtube. Content creator biasanya akan mengulas produk, kemudian para viewer akan memberikan komentar terkait produk tersebut. Disini, peneliti akan menggunakan metode Latent Dirichlet Allocation (LDA) untuk mempercepat proses penyerapan informasi atas produk tersebut. Sehingga calon konsumen lebih mudah mengidentifikasi produk yang akan dibeli berdasarkan ulasan content creator dan viewer sebelumnya berdasarkan topik yang teridentifikasi dari klasterisasi. Hasil LDA ini juga dapat digunakan oleh pemilik brand produk tersebut untuk melihat tingkat acceptance masyarakat terhadap produknya. Hasil akhir dari penelitian ini adalah perancangan model LDA untuk topik pelanggan.

Kata kunci: topic clustering; review; Latent Dirichlet Allocation; Youtube

1. Pendahuluan

Pemodelan topik adalah teknik pembelajaran mesin yang digunakan di berbagai domain, termasuk rekayasa perangkat lunak, penambangan teks, dan perawatan kesehatan, untuk mengungkap struktur semantik yang mendasari kumpulan dokumen (1,2). Secara otomatis mendeteksi topik dengan menganalisis pola dan hubungan antar kata dalam teks (1). Pemodelan topik telah diterapkan dalam penelitian rekayasa perangkat lunak untuk menganalisis data tekstual, mendukung pemahaman kode sumber, dan mengidentifikasi apa yang didiskusikan pengembang secara online (3). Dalam kesehatan, pemodelan topik telah digunakan untuk mengklasifikasikan penyakit, seperti pseudogout, menggunakan data catatan kesehatan elektronik (4). Pemodelan topik juga telah diterapkan untuk menganalisis literatur ilmiah tentang topik-topik seperti coronavirus (5). Teknik pemodelan topik meliputi Latent

Dirichlet Allocation (LDA), analisis semantik laten (LSA), dan analisis semantik laten probabilistik (PLSA) (6). Secara keseluruhan, pemodelan topik adalah alat yang berharga untuk mengekstraksi informasi yang bermakna dan mendapatkan wawasan dari kumpulan besar dokumen teks di berbagai domain.

Pemodelan topik berharga dalam konteks produksi dan penjualan produk karena beberapa alasan. **Pertama**, dapat membantu pemilik bisnis memahami karakteristik terpenting dari suatu produk atau jasa di mata pelanggan (7). Dengan menganalisis ulasan online menggunakan pemodelan topik, manajer dapat mengidentifikasi pola dan wawasan utama yang dapat menginformasikan pengembangan produk dan strategi pemasaran. **Kedua**, pemodelan topik dapat digunakan untuk mengurangi dimensi ulasan online, terutama karena volume ulasan meningkat seiring dengan perluasan penjualan online (7). Ini memungkinkan manajer untuk mengekstrak informasi yang paling relevan dan bermakna dari sejumlah besar umpan balik pelanggan, memungkinkan mereka membuat keputusan berdasarkan data.

Selanjutnya, pemodelan topik telah diterapkan di berbagai domain yang terkait dengan penjualan produk, seperti mempelajari dampak komentar online terhadap penjualan produk (8), menganalisis gaya persuasif linguistik di lingkungan e-commerce sosial (9), dan mengidentifikasi atribut produk penting untuk prediksi penjualan [5]. Aplikasi ini menunjukkan keserbagunaan dan keefektifan pemodelan topik dalam memahami perilaku konsumen dan mendorong penjualan.

Singkatnya, pemodelan topik adalah alat yang berharga dalam konteks produksi dan penjualan produk karena memungkinkan bisnis memperoleh wawasan dari umpan balik pelanggan, mengurangi dimensi ulasan online, dan membuat keputusan berdasarkan data untuk meningkatkan pengembangan produk dan strategi pemasaran.

Latent Dirichlet Allocation (LDA) adalah teknik pemodelan topik yang banyak digunakan dalam pemrosesan bahasa alami (NLP) yang secara otomatis mendeteksi topik dalam kumpulan besar dokumen teks George & Sumathy (1). LDA mengasumsikan bahwa setiap dokumen adalah campuran topik, dan setiap topik adalah distribusi kata (10). Ini didasarkan pada distribusi Dirichlet, yang merupakan distribusi probabilitas atas distribusi multinomial (11).

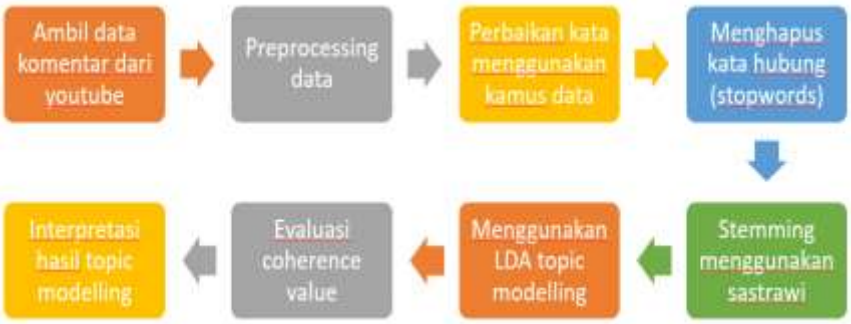
Beberapa peneliti telah memanfaatkan Alokasi Dirichlet Laten (LDA) dalam studi pemodelan topik mereka. mempekerjakan LDA untuk menemukan topik laten dalam tinjauan literatur tentang realitas virtual dalam pemasaran (12). menggunakan LDA, bersama dengan perpustakaan pyLDAvis, untuk melakukan pemodelan topik dan mengkarakterisasi topik laten di korpus area penelitian (13). menggunakan LDA untuk pemodelan topik dalam ulasan mereka tentang penelitian kecerdasan buatan dalam domain klinis (14). menyoroti popularitas LDA di bidang penambangan data untuk pemodelan topik di forum penjawab pertanyaan komunitas (15).

Secara keseluruhan, LDA adalah teknik yang ampuh untuk mengungkap topik laten dalam data tekstual yang tidak terstruktur dan telah digunakan secara luas dalam penelitian di berbagai domain.

Dari deskripsi diatas, maka peneliti tertarik untuk menerapkan topic analysis menggunakan LDA dengan studi kasus komentar dari review produk oleh influencer.

2. Metode

Metode penelitian yang peneliti usulkan dalam pemanfaatan topic modelling berbasis LDA dapat digambarkan sebagai berikut:

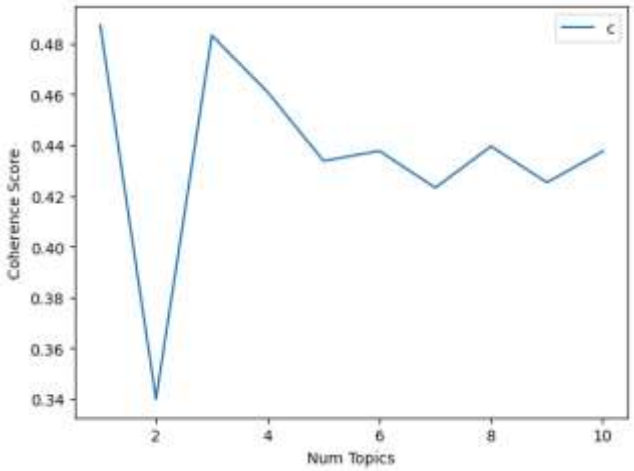


Gambar 1 metode yang diusulkan

Penjelasan pada gambar 1 sebagai berikut: Pertama, peneliti mengambil data komentar dari channel youtube seorang influencer. Untuk mengambil data komentar, peneliti menggunakan tools dari https://communalytic.org/. Setelah data berhasil diekstrak, tahapan kedua adalah melakukan preprocessing, antara lain: menghapus angka, menghapus emoji, menghapus url, menghapus tanda baca, dan membuat karakter menjadi huruf kecil (lowecase). Tahapan ketiga menggunakan kamus perbaikan kata digunakan untuk merubah kata-kata menjadi kata baku. Tahapan keempat adalah menghapus kata sambung (stopwords). Tahapan kelima adalah stemming menggunakan sastrawi untuk mendapatkan kata dasar. Tahapan keenam menggunakan LDA sebagai topic modelling. Tahapan ketujuh adalah evaluasi topic modelling menggunakan coherence value. Tahapan kedelapan adalah pembacaan hasil topic modelling tersebut.

3. Hasil dan Pembahasan

Realisasi dari metode yang diusulkan adalah sebagai berikut: pertama, peneliti mengambil data komentar review dari channel gadgetin dengan kanal https://www.youtube.com/watch?v=PCs-8KTlhKg yang membahas produk smartphone merk tertentu. Data ditarik menggunakan tools communalytic. Saat penarikan, 2193 record berhasil ditarik. Kedua, data tersebut dilakukan preprocessing (menghapus angka, menghapus emoji, menghapus url, menghapus tanda baca, dan membuat karakter menjadi huruf kecil). Ketiga, data yang sudah dipreprocessing, dinormalisasi datanya agar menjadi kata baku. Dikarenakan sudah menjadi kata baku, maka tahapan keempat adalah menghapus kata sambung (stopwords). Setelah kata sambung berhasil dihilangkan, maka tahapan kelima adalah melakukan stemming untuk mendapatkan kata dasar. Tahapan keenam menggunakan LDA sebagai topic modelling. Disini peneliti menggunakan library pyldavis==3.2.1. Tahapan ketujuh, peneliti menggunakan coherence value untuk mendapatkan jumlah topik optimal sesuai gambar 2:



Gambar 2 nilai coherence value berdasarkan jumlah topik

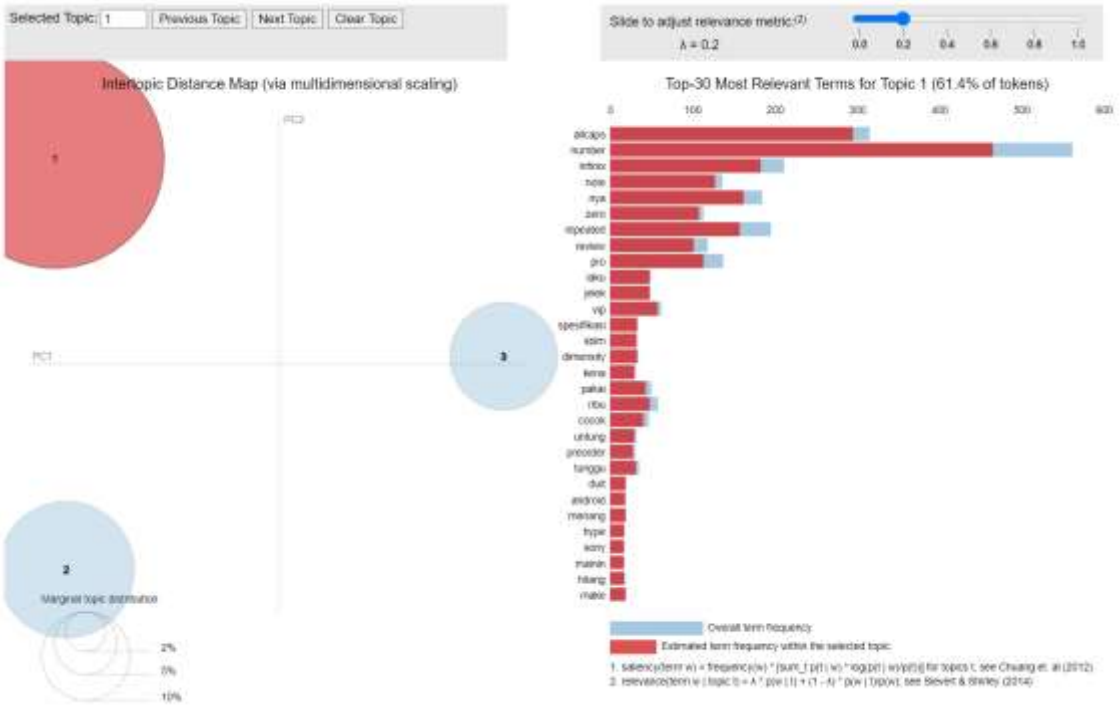
Pada gambar 2, dapat dilihat topik optimal berada pada angka 3. Maka dari itu, peneliti memilih 3 topic untuk melihat hasil topic clustering nya. Untuk populasi cluster dapat dilihat pada tabel 1:

Tabel 1 hasil topic clustering

Topic 1	0.029*"bang" + 0.023*"number" + 0.021*"ghoib" + 0.020*"hp" + 0.016*"user" + 0.015*"repeated" + 0.015*"harga" + 0.015*"official" + 0.014*"sih" + 0.013*"store"
Topic 2	0.082*"number" + 0.052*"allcaps" + 0.032*"infinix" + 0.028*"nya" + 0.028*"repeated" + 0.022*"note" + 0.020*"pro" + 0.019*"bang" + 0.019*"zero" + 0.018*"review"
Topic 3	0.033*"number" + 0.033*"user" + 0.024*"harga" + 0.018*"tengkulak_borong" + 0.018*"mati_masyarakat" + 0.018*"paham_mafia" + 0.015*"bang" + 0.014*"priantoko" + 0.012*"kakak" + 0.012*"barang"

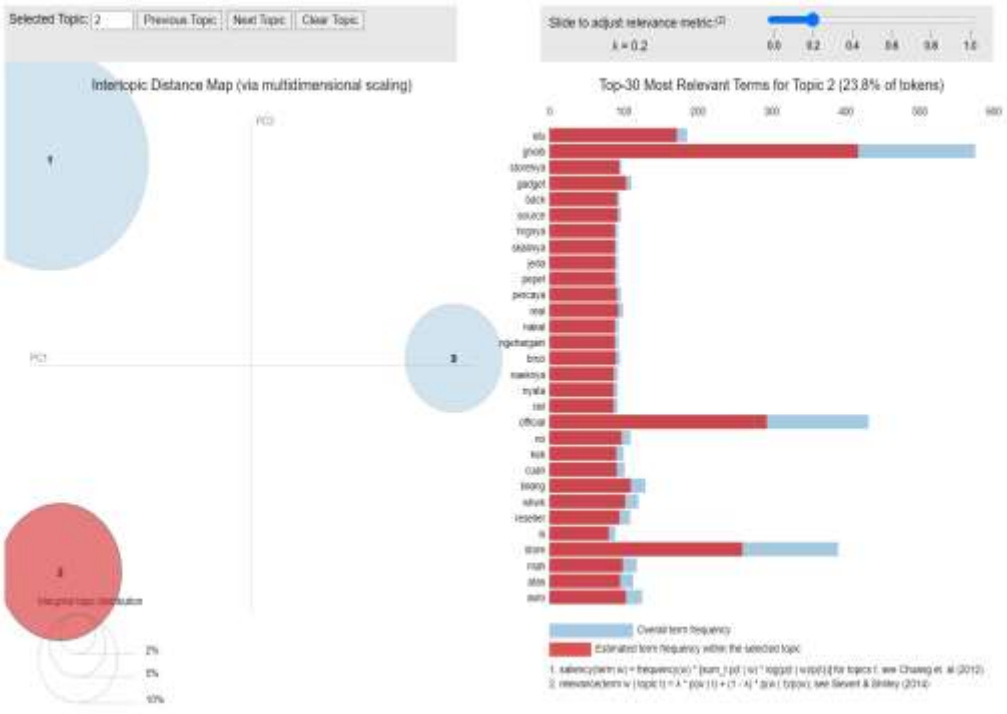
Pada tabel 1, ada beberapa kata yang muncul karena penggunaan preprocessing, terutama pada bagian menghapus angka, disini angka akan direname menjadi number.

Pada tahapan kedelapan, peneliti memvisualisasikan hasil topic clustering yang didapatkan sesuai pada gambar 3,4 dan 5 dengan relevance matric sebesar 0.2 sebagai berikut.



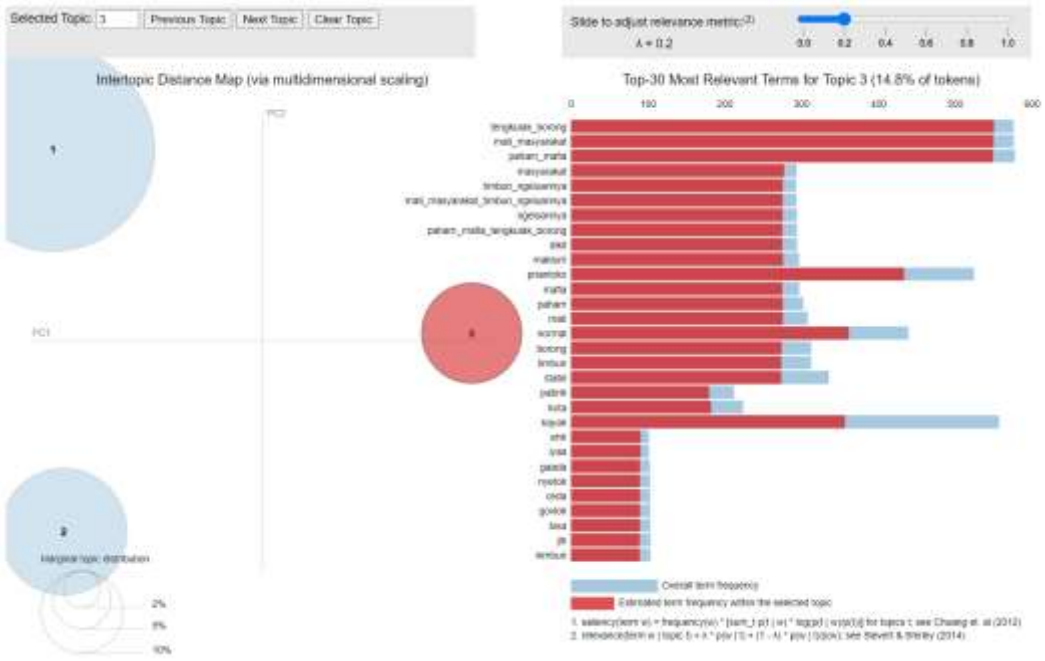
Gambar 3 hasil visualisasi topic 1

Pada gambar 3, topic 1 mempunyai hasil topic yang membahas tentang spesifikasi dan kualitas.



Gambar 4 hasil topic 2

Pada gambar 4, topic 2 berkaitan dengan harga dan ketersediaan produk.



Gambar 5 hasil topic 3

Pada gambar 5, topic 3 berkaitan dengan penimbunan produk tersebut.

4. Kesimpulan

Berdasarkan pembahasan pada subbab sebelumnya, dapat diambil kesimpulan bahwa Pengembangan Model Analisa Diskusi Topic Produk Menggunakan Latent Dirichlet Allocation Pada Platform Youtube berhasil dilaksanakan.

Meskipun demikian, penelitian selanjutnya, peneliti akan memproses beberapa pendekatan, antara lain: 1) mencoba menghapus kata-kata yang terjadi dikarenakan hasil preprocessing; 2) tidak menggunakan stop words; 3) tidak menggunakan kamus data.

5. Referensi

[1]. Lijimol George, P. Sumathy. An Integrated Clustering and BERT Framework for Improved Topic Modeling. Int J Inf Technol. 2023;

[2]. Pooja Kherwa, Poonam Bansal. Topic Modeling: A Comprehensive Review. Icst Trans Scalable Inf Syst. 2018;

[3]. Camila Daiane Silva, Matthias Galster, Fabian Gilson. Topic Modeling in Software Engineering Research. Empir Softw Eng. 2021;

[4]. Sara K. Tedeschi, Tianxi Cai, Zeling He, Yuri Ahuja, Chuan Hong, Katherine P. Yates, et al. Classifying Pseudogout Using Machine Learning Approaches With Electronic Health Record Data. Arthritis Care Res. 2021;

[5]. Barkha Bansal, Sangeet Srivastava. Context-Sensitive and Attribute-Based Sentiment Classification of Online Consumer-Generated Content. Kybernetes. 2019;

[6]. Wonkwang Jo, Yeol Kim, Minji Seo, Nayoung Lee, Junli Park. Online Information Analysis on Pancreatic Cancer in Korea Using Structural Topic Model. Sci Rep. 2022;

[7]. Atieh Poushneh, Reza Rajabi. Can Reviews Predict Reviewers' Numerical Ratings? The Underlying Mechanisms of Customers' Decisions to Rate Products Using Latent Dirichlet Allocation (LDA). J Consum Mark. 2022;

[8]. Yongjie Chu, Jingjing Cao. Exploring the Influence of E-Commerce Multi-Modal Data on Online Shopping Based on Stepwise Regression Model. 2023;

[9]. Hanyang Luo, Sijia Cheng, Wanhua Zhou, Su-min Yu, Xudong Lin. A Study on the Impact of Linguistic Persuasive Styles on the Sales Volume of Live Streaming Products in Social E-Commerce Environment. Mathematics. 2021;

[10]. Mohit Garg, Priya Rangra. Bibliometric Analysis of Latent Dirichlet Allocation. Desidoc J Libr Inf Technol. 2022;

[11]. Amritanshu Agrawal, Wei Fu, Tim Menzies. What Is Wrong With Topic Modeling? And How to Fix It Using Search-Based Software Engineering. Inf Softw Technol. 2018;

[12]. Sandra Loureiro, João Farias Guerreiro, Sara Eloy, Daniela Langaro, Padma Panchapakesan. Understanding the Use of Virtual Reality in Marketing: A Text Mining-Based Review. J Bus Res. 2019;

[13]. Tristan Lim. Environmental, Social, and Governance (ESG) and Artificial Intelligence in Finance: State-of-the-Art and Research Takeaways. 2023;

[14]. Renu Sabharwal, Shah Jahan Miah, Samuel Fosso Wamba. Extending Artificial Intelligence Research in the Clinical Domain: A Theoretical Perspective. Ann Oper Res. 2022;

[15]. Muhammad Usman, Farwa Ahmad, Usman Habib, Adeel Ashraf Cheema, Zhiguang Qin, Muhammad Ahmad. Combining Latent Factor Model for Dynamic Recommendations in Community Question Answering Forums. Comput Intell Neurosci. 2022;