Seminar Nasional Sains Data 2024 (SENADA 2024)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Analisis Media Sosial dan Massa Terhadap Industri Otomotif Indonesia Menggunakan Semi-supervised LDA.

Alfathamdi Putra Umaryadi[1]
[1]*Matematika, Universitas Islam Negeri Syarif Hidayatullah Jakarta*
[1]alfathamdi2@gmail.com

Abstract: *Every year, the automotive industry in Indonesia produces more than 1 million units, with export contributions amounting to USD 258.82 billion in 2024. Mass media and social media are sources of public information that can be utilized to enhance the growth of the automotive industry. However, discussions on social media and news in mass media often lack direction due to the wide range of topics discussed, making analysis difficult. This research aims to classify topics into 5 categories, namely Cost, Features, Reliability, Security, Safety, and Environmental Friendliness. The semi-supervised LDA topic modeling method is used to classify data sourced from 15 mass media platforms such as Otomotif Kompas, Gaikindo, Carmudi Indonesia, and others, as well as 5 social media platforms such as TikTok, X (Twitter), Facebook, and others. Evaluation of the model created shows a coherence level of 2.5. Based on this research, using a relevance level of 0.2, researchers found that in terms of Cost, the Honda brand is the most frequently discussed, while the Daihatsu brand is the most discussed in terms of Environmental Friendliness, and Toyota is the most discussed brand in terms of Reliability.*
Keywords: Automotive Industry, Mass Media, Semi-supervised LDA, Social Media, Topic Analysis

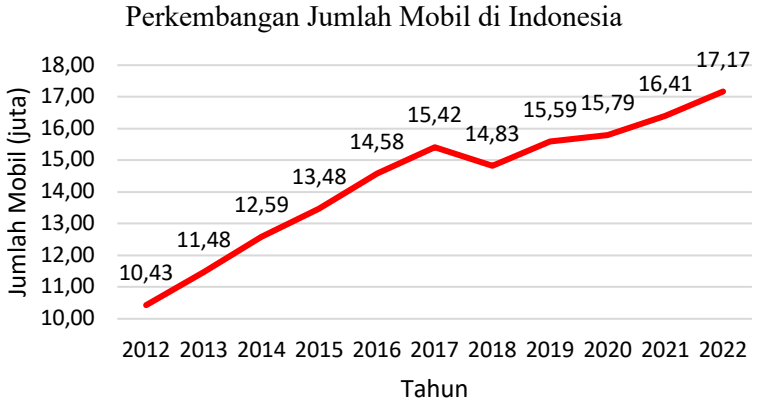
Abstrak: Setiap tahunnya, industri otomotif di Indonesia memproduksi lebih dari 1 juta unit, dan kontribusi ekspor sebesar USD 258,82 Miliar pada tahun 2024. Media massa dan media sosial sumber informasi publik yang dapat digunakan untuk meningkatkan pertumbuhan industri otomotif. Namun, perbincangan di media sosial dan berita di media massa sering kali tidak terarah karena beragamnya topik yang dibahas, sehingga sulit untuk melakukan analisis. Penelitian ini bertujuan untuk mengklasifikasikan topik-topik menjadi 5 kategori, yaitu Biaya, Fitur, Reliabilitas, Keamanan, Keselamatan, Ramah Lingkungan, dan Lainnya. Metode topik modeling semi-supervised LDA digunakan untuk mengklasifikasi data yang bersumber dari 15 platform media massa seperti Otomotif Kompas, Gaikindo, Carmudi Indonesia, dan lainnya, serta 5 platform media sosial seperti TikTok, X (Twitter), Facebook, dan lainnya. Evaluasi terhadap model yang dibuat menunjukkan tingkat koherensi sebesar 2,5. Berdasarkan penelitian ini, dengan menggunakan tingkat relevansi sebesar 0,2, peneliti menemukan bahwa pada aspek Biaya, merek Honda yang paling sering diperbincangkan, sementara merek Daihatsu menjadi yang paling diperbincangkan dalam aspek Ramah Lingkungan, dan Toyota adalah merek yang paling diperbincangkan dalam aspek Reliabilitas.
Kata kunci: Analisis Topik, Industri Otomotif, LDA Semi-supervised, Media Massa, Media Sosial

I. PENDAHULUAN

Dari gambar 1.1 diperlihatkan jika dalam 10 tahun, negara Indonesia mengalami pertumbuhan produksi mobil dengan angka yang sangat signifikan, yaitu sebesar 64,6%. Fenomena ini menunjukan bahwa transportasi merupakan faktor penting untuk meningkatkan kualitas hidup [1]. Upaya untuk menyediakan layanan dan sistem transportasi yang efisien, tidak hanya akan memberikan dampak signifikan bagi konsumen, tapi juiyaaga bagi berpotensi untuk meningkatkan pangsa pasar perusahaan otomotif. Sumber informasi dari media massa dan media sosial kini sangatlah beragam. Media massa tidak hanya memberikan wawasan mengenai urusan publik, namun juga membantu para pembacanya juga belajar mengenai berbagai hal berdasarkan penekanan topik yang terdapat dalam berita [2]. Di sisi lain, media sosial sering kali berfungsi sebagai wadah opini publik. Dengan pengguna Media sosial di indonesia lebih dari 139 juta jiwa, atau 49.9% dari total populasi[1], media sosial memainkan peran krusial dalam membentuk persepsi terhadap berbagai industri, termasuk industri otomotif.

[1] M. A. Rizaty, "Data Jumlah Pengguna Media Sosial di Indonesia pada 2024," *dataindonesia.id*, 4 April 2024, *https://dataindonesia.id/internet/detail/data-jumlah-pengguna-media-sosial-di-indonesia-pada-2024*

E-ISSN 2808-5841
P-ISSN 2808-7283



Gambar 1. Pertumbuhan Produksi Mobil di Indonesia[2].

Pengaruh media massa dan media sosial terhadap industri otomotif sangatlah signifikan. Persepsi dan tren yang terbentuk dari berbagai platform ini dapat mempengaruhi perkembangan industri secara keseluruhan. Namun, dengan pertumbuhan informasi yang eksponensial di internet, tantangan muncul dalam menemukan sumber informasi yang konsisten dan relevan. Informasi yang tersebar di media sosial, terutama dalam bentuk hoaks, juga menambah kompleksitas pencarian informasi yang relevan [3,5]. Faktor lain yang membuat sulitnya menemukan informasi yang relevan adalah keberadaan berbagai platform alternatif media sosial dan situs web media massa yang dapat digunakan. Hal ini menyebabkan topik informasi yang ada atau yang dicari berpotensi tersembunyi di platform media lain [6]. Selain itu, gaya penulisan yang digunakan oleh individu berbeda-beda, dan setiap institusi memiliki standar masing-masing dalam mempublikasikan berita mereka. Variasi gaya penulisan ini dapat menurunkan akurasi dan meningkatkan kesulitan dalam menginterpretasikan isu-isu dan tren terkait topik dalam informasi yang beredar.

Dengan pertumbuhan informasi yang eksponensial setiap detiknya, dengan keberagaman data yang makin meningkat, dan hanya sedikit data yang yang memiliki label. Oleh karena itu diperlukan pengkategorian berdasarkan topik dari artikel atau opini dari data, dengan bantuan klasifikasi *semi-supervised.* Memperoleh data pelatihan yang berlabel dari dataset besar memerlukan waktu dan bantuan manusia, sehingga penting untuk mengimplementasikan mekanisme pemilihan data yang informatif selama pengembangan model demi hasil kinerja optimal, dengan mempertimbangkan pendekatan *active learning* dalam kerangka kecerdasan buatan berbasis data [7]. Tantangan utama dalam melakukan analisis topik media massa dan media sosial adalah keberagaman data yang tidak terkategorikan dengan baik, yang menyebabkan informasi menjadi kurang konsisten dan kurang relevan untuk memahami isu-isu, tren, dan pandangan sosial terhadap industri otomotif. *Semi-supervised LDA*, menjadi solusi yang potensial untuk mengatasi tantangan ini. Dengan memanfaatkan data yang heterogen dan tidak terstruktur, pendekatan ini dapat membantu mengidentifikasi dan mengkategorikan informasi yang belum berlabel dengan baik. Meskipun LDA pada dasarnya adalah algoritma *unsupervised*, pendekatan *semi-supervised* memungkinkan untuk inisiasi label pada tahap awal penelitian, memperbaiki akurasi dalam klasifikasi topik [8,10]. Terdapat beberapa pendekatan *semi-supervised* lain, seperti menggunakan framework *Variational Autoencoder* (VAE), dan metode yang fokus pada klasifikasi

[2] A. Ahdiat, "Ini Pertumbuhan Jumlah Mobil di Indonesia 10 Tahun Terakhir," *databoks*, 15 Maret, 2023, https://databoks.katadata.co.id/datapublish/2023/03/15/ini-pertumbuhan-jumlah-mobil-di-indonesia-10-tahun-terakhir

dokumen [11]. Namun, penggunaan *VAE framework* untuk melakukan pembelajar semi-supervised masih jarang di dilakukan [12], disini kita mengajukan *Neural Labeled LDA*. Dalam konteks industri otomotif, pemahaman yang lebih baik terhadap isu-isu, tren, dan pandangan sosial yang berkembang melalui media massa dan media sosial menjadi krusial. Model seperti *Neural Labeled LDA* dapat menjadi alternatif yang menarik untuk memperkuat analisis topik. model yang diusulkan memberikan model pengetahuan awal kepada model sebelum melanjutkan ke tahap topik modeling, untuk menghasilkan model yang lebih terarah dalam melakukan klasifikasi topik.

Pada penelitian sebelumnya sudah pernah menggunakan pendekatan *unsupervised LDA*, klasifikasi topik dilakukan berdasarkan distribusi topik dalam dokumen, negara/wilayah, dan waktu dari data riset transportasi [1]. Kelebihan dari penelitian yang dilakukan sebelumnya adalah, penggunaan *wordcloud* untuk mempermudah hasil interpretasi, sehingga isi dari setiap topik dapat dipahami lebih baik. Namun kelemahan pada jurnal ini adalah, model LDA menghasilkan lebih dari 50 topik yang berbeda, dan label topik yang ditetapkan berdasarkan distribusi kata tanpa konsultasi dengan ahlinya. Selain itu pendekatan *unsupervised LDA* dapat menyebabkan interpretasi yang bervariasi dari topik yang sama. serta kurang efektifnya dalam mengklasifikasi topik-topik yang benar-benar baru atau tidak terduga sebagaimana yang dapat dilakukan oleh *unsupervised LDA*. Untuk mengatasi tantangan tersebut, kami mengusulkan pendekatan *semi-supervised LDA*. *semi-supervised LDA* dapat meningkatkan kemampuan model dalam melakukan pengelompokan topik topik yang muncul dengan lebih konsisten dan mengidentifikasi tren yang belum ada. Pendekatan ini memberikan arahan yang lebih terarah dalam analisis topik, meningkatkan kemungkinan interpretasi isu isu lebih relevan.

II. METODE PENELITIAN

II.1 Pengumpulan Data

Untuk memahami secara menyeluruh persepsi merek dan dinamika pasar dalam konteks regulasi ketat terhadap data media sosial dan berita media massa. makalah kami memperkenalkan Analitik Berbasis Relevansi. Metodologi inovatif ini terdiri dari dua komponen utama: Pengambilan Data yang Relevan dan Pembelajaran Mesin Berbasis Pengambilan Data yang Relevan. Setiap komponen dirancang dengan hati-hati untuk mengatasi tantangan etis dalam pengumpulan data sambil memastikan analisis yang mendalam terhadap konten yang dihasilkan oleh pengguna dan berita yang dihasilkan oleh institusi.

II.2 Strategi Pengumpulan Data

Untuk menghadapi tantangan yang disebabkan oleh kekhawatiran privasi dan batasan regulasi, kami mengadopsi metode Pengambilan Data Relevan. Metodologi ini melibatkan penggunaan data yang telah diindeks oleh mesin pencari untuk membuat sampel representatif dari konten media sosial dan berita media massa yang terkait dengan kehadiran merek otomotif internasional di Indonesia. Dengan berfokus pada data yang tersedia melalui indeks mesin pencari publik, kami menjamin kepatuhan terhadap kebijakan perlindungan data sambil mempertahankan kekayaan dan keragaman dataset.

II.3 Integrasi API Protokol Hukum

Untuk mematuhi praktik penelitian yang etis dan meningkatkan keandalan pengumpulan data kami, kami mengintegrasikan API protokol hukum ke dalam proses Pengambilan Data Relevan. API ini memfasilitasi ekstraksi data sesuai dengan pedoman hukum yang berlaku, menambahkan lapisan jaminan tambahan bahwa metodologi kami selaras dengan kerangka kerja regulasi yang ada. Integrasi

ini memastikan bahwa pengumpulan data kami tidak hanya memenuhi standar privasi, tetapi juga beroperasi sesuai dengan batasan protokol hukum.

II.4 Kriteria Pengambilan Data

Istilah "Relevan" dalam metodologi kami mencerminkan pendekatan strategis terhadap pengambilan data. Kami menentukan relevansi berdasarkan kriteria yang sudah ditetapkan sebelumnya, seperti konten yang berhubungan dengan merek otomotif, berasal dari pengguna di Indonesia, dan mencakup periode waktu tertentu. Pendekatan pengambilan sampel yang terfokus ini memungkinkan kami untuk mengumpulkan data yang tidak hanya mematuhi regulasi privasi tetapi juga relevan dengan tujuan penelitian kami, sehingga memastikan analisis yang tepat dan bermakna.

II.5 Kata Kunci Pengambilan Sampel

Kata kunci yang kami gunakan dalam pengambilan sampel melibatkan istilah-istilah yang memastikan data yang terambil adalah data dalam bahasa Indonesia, seperti:

- toyota indonesia
- daihatsu indonesia
- honda indonesia
- hyundai indonesia
- suzuki indonesia
- mitsubishi indonesia
- wuling indonesia
- mobil toyota
- mobil daihatsu
- mobil honda
- mobil hyundai
- mobil suzuki
- mobil mitsubishi
- mobil wuling
- dll

II.6 *Data Cleaning and Preprocessing*

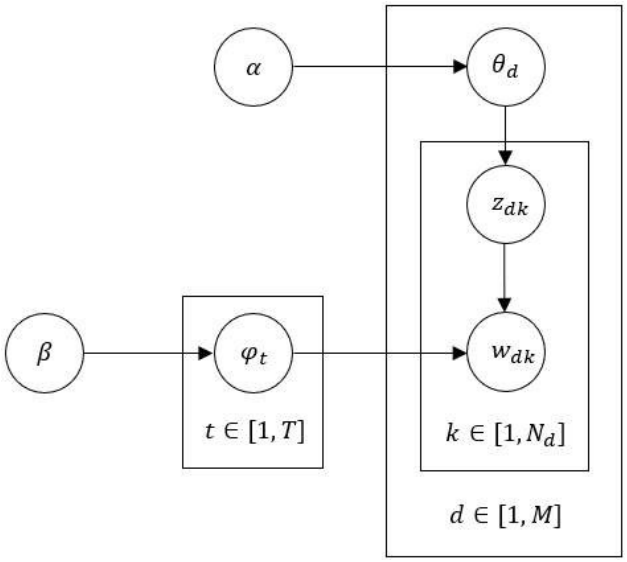
Data sampel yang sudah dikumpulkan sebelumnya merupakan data yang berisi opini atau konten dari media sosial dan berita dari media massa, oleh karena itu data yang sudah dikumpulkan akan diproses (*pre-processing)* terlebih dahulu untuk menghilangkan hal yang tidak relevan, atau menurunkan hal yang tidak penting, sehingga interpretasi yang didapatkan nantinya lebih relevan. Di tahap ini akan dilakukan *Tokenisasi, lowercase,* penghilangan *non-alphabet character, stopwords, punctuation,* dan *lemmatization* terhadap data sampel.

II.7 *Vectorization.*

Bag of Words adalah salah satu teknik untuk mempresentasikan data kedalam bentuk numerik, salah satu algoritma yang sering digunakan adalah *Term Frequency,* yaitu algoritma untuk menghitung bobot dari setiap data di dokumen yang dimiliki. Bobot pada *Term Frequency* adalah nilai yang menunjukan banyaknya suatu data muncul dalam dataset yang dimiliki.

II.8 *Latent Dirichlet Allocation*

Latent Dirichlet Allocation (LDA) merupakan salah satu algoritma untuk mengklasifikasikan data dari dokumen. LDA dapat mengelompokan data berdasarkan bobot dari setiap kata dalam dokumen, dan mengelompokkannya ke dalam berbagai topik. Sebagai model *unsupervised*, LDA tidak memerlukan label. Semua topik yang ditentukan oleh LDA merupakan proses perhitungan statistik terhadap kata-kata dalam dokumen.



Gambar 2. Grafik Representasi Model LDA bersumber dari Sun, Lijun [1].

Tabel 1. Notasi Variabel dan Parameter

Notasi	Keterangan
t	Index topic
d	Index dokumen
k	Index Kata
α	*Dirichlet prior* pada distribusi topik per dokumen *(hyperparameter)*
β	*Dirichlet prior* pada distribusi kata per topik *(hyperparameter)*
θ_d	Distribusi topik dari dokumen d
φ_t	Distribusi kata dari topik t
w_{dk}	Kata k dalam koleksi kata dari dokumen d
z_{dk}	Topik yang ditentukan dari kata w_{dk}
T	Banyak Topik
M	Banyak dokumen
V	Banyak kata dalam kosakata
N_d	Banyak kata dalam dokumen d

Gambar 2. adalah tahapan bagaimana LDA bekerja dalam bentuk notasi. Pada tahap awal LDA mendefinisikan T topik, dan setiap topik t dikaitkan dengan dengan distribusi kata dari topik t terhadap kata dalam kosakata. Distribusi kata dari topik t dipilih dari distribusi Dirichlet, Dirichlet $V(\beta)$. Dari topik yang telah ditentukan sebelumnya, dokumen d dihasilkan dari sampling distribusi θ_d atas T topik dari distribusi Dirichlet lain, Dirichlet $T(\alpha)$, yang menentukan topik untuk setiap kata dalam koleksi kata dari dokumen d, dan kemudian memilih setiap kata w_{dk} berdasarkan θ_d. Dalam menghasilkan setiap kata w_{dk}, LDA mengambil sampel dari topik tertentu $z_{dk} \in [1, K]$ dari distribusi multinomial, Multinomial $K(\theta_d)$, Kemudian, kata w_{dk} dipilih dari distribusi multinomial, Multinomial $V(\varphi_{z_{dk}})$.

II.9 *Semi-supervised Topic Modelling*

LDA merupakan model *unsupervised*, sehingga tidak memerlukan pengetahuan tentang jumlah topik untuk melakukan *topik modelling*. Dikarenakan topik yang diperoleh merupakan hasil struktur statistik dari matriks dokumen-kata berdasarkan likelihood maksimal atau korelasi prinsipal maksimal. Namun dengan memberikan pengetahuan di awal, sebelum lanjut ke tahap proses *topic modelling*, model unsupervised menjadi model *semi-supervised*, hal ini dilakukan agar kualitas interpretasi topik yang diperoleh lebih tinggi, serta klasifikasi yang diperoleh lebih relevan.

II.10 *Validation*

Performa dari *semi-supervised LDA* akan dihitung berdasarkan *coherence score*-nya, *coherence* merupakan salah satu evaluasi untuk menilai performa topik model. *Coherence* digunakan untuk menganalisis hubungan dari dua dataset yang mirip. Dalam perhitungan *coherence topik modeling,* kualitas data dihitung. Dalam penelitian kualitas data akan evaluasi dengan menghitung *coherence score*-nya.

$$\vec{v}(W') = \left\{ \sum_{w_i \in W'} NPMI(w_i, w_j)^{\gamma} \right\}_{j=1,\dots,|W|} \tag{1}$$

$$NPMI(w_i, w_j)^{\gamma} = \left(\frac{log \frac{P(w_i, w_j)+\epsilon}{P(w_i) \cdot P(w_j)}}{-log(P(w_i, w_j)+\epsilon)} \right)^{\gamma}, \tag{2}$$

$$\phi_{S_i}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \, w_i}{\|\vec{u}\|_2 \cdot \|\vec{w}\|_2} \tag{3}$$

Pada persamaan (1), $\vec{v}(W')$ adalah vektor untuk himpunan W'. $\sum_{w_i \in W'}$ merupakan agregasi setiap w_i dalam himpunan W'. Pada persamaan (2), NPMI (Normalized Pointwise Mutual Information) antara kata $w_i, dan\ w_j$ digunakan untuk mengukur keterkaitan antar kedua kata dengan menggunakan logaritma dari rasio probabilitas gabungan dengan ϵ untuk menghindari terjadinya logaritma nol. γ memberikan bobot kepada nilai NPMI yang lebih tinggi. Pada persamaan (3) didefinisikan $\phi_{S_i}(\vec{u}, \vec{w})$ sebagai kesamaan kosinus antara kedua vektor atau disebut sebagai *coherence score*, yang menunjukan seberapa kuat kata-kata dalam himpunan W, Mendukung W'. $\phi_{S_i}(\vec{u}, \vec{w})$ memiliki range nilai dari 0 hingga 1, yang dimana semakin tinggi nilai *coherence*-nya maka akan semakin baik generalisasi topik yang dibentuk. Berikut adalah kriteria nilai *coherence*-nya.

Tabel 2. Kriteria Nilai Koherensi

Rentang nilai koherensi	Keterangan
0 - 0.29	Rendah
0.3 - 0.59	Cukup
0.6 - 1	Baik

II.11 Analisis Hasil

Analisis akhir dalam penelitian ini melibatkan analisis tren topik di industri otomotif menggunakan *semi-supervised LDA*, visualisasi *wordcloud* dan analisis frekuensi kata dari setiap media untuk

memudahkan pemahaman distribusi kata dalam setiap topik, serta penelitian mendalam terhadap masing-masing topik untuk mendapatkan insight yang lebih mendetail.

III. HASIL DAN PEMBAHASAN

Di bagian pertama akan dipresentasikan terlebih dahulu *Unigram, Bigram*, dan *Trigram* yang sering muncul, baik secara umum, dalam media sosial, dan media massa, beserta visualisasi *wordcloud*-nya. Kemudian akan dilakukan analisis terhadap model. Lalu terkahir, akan dilakukan analisis tren dari setiap topiknya.

III.1 *Unigram, Bigram*, dan *Trigram*

Gambar 3a. Wordcloud Media Sosial dari Biaya

Gambar 3b. Wordcloud Media Sosial dari Fitur

Gambar 3c. Wordcloud Media Sosial dari Keselamatan

Gambar 3d. Wordcloud Media Sosial dari Lingkungan

Gambar 3e. Wordcloud Media Sosial dari Reliabilitas

Gambar 3f. Wordcloud Media Sosial dari Lainnya

E-ISSN 2808-5841
P-ISSN 2808-7283

Tabel 3. Daftar Unigram, Bigram, dan Trigram yang dipilih dalam media sosial

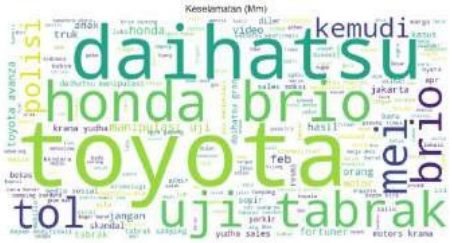
Kalimat	Kemunculan	Kalimat	Kemunculan	Kalimat	Kemunculan
Unigram		**Bigram**		**Trigram**	
Toyota	2.29%	Honda Brio	0.24%	Toyota Kijang Innova	0.0.7%
Honda	2.12%	Toyota Avanza	0.13%	Daihatsu Sahabat Daihatsu	0.04%
Daihatsu	1.29%	Toyota Toyota	0.12%	Harga Harga Bekas	0.03%
Harga	0.77%	Daihatsu Daihatsuin	0.12%	Harga Bekas Tarik	0.03%
Baru	0.73%	Toyota Innova	0.11%	Toko Velg Ban	0.03%
Tahun	0.73%	Toyota Fortuner	0.11%	Pasang Kaca Film	0.03%
Kendara	0.58%	Toyota Kijang	0.11%	Salesman Salesgirl Usaha	0.03%
Info	0.55%	Kaca Film	0.11%	Harga Harga Toyota	0.02%
Promo	0.54%	Sahabat Daihatsu	0.10%	Bekas Toyota Kijang	0.02%
Motor	0.45%	Honda Jaz	0.10%	Harga Bekas Toyota	0.02%

Dalam media sosial, topik biaya seringkali mengangkat diskusi seputar promo serta merek Honda dan Toyota. Fitur kendaraan seperti kaca film juga menjadi fokus pembicaraan, sering kali terkait dengan merek Toyota dan Honda. Dalam konteks keselamatan, kata-kata seperti korban, warga, orang, dan kendaraan muncul secara signifikan. Ketika membahas lingkungan, isu terkait kata kerja dan merek Toyota serta Honda sering menjadi perhatian utama. Pada topik reliabilitas, mesin dan merek Toyota sering dibahas secara mendalam.

Gambar 3g. Wordcloud Media Massa dari Biaya

Gambar 3h. Wordcloud Media Massa dari Fitur

E-ISSN 2808-5841
P-ISSN 2808-7283



Gambar 3i. Wordcloud Media Massa dari Keselamatan

Gambar 3j. Wordcloud Media Massa dari Lingkungan

Gambar 3k. Wordcloud Media Massa dari Reliabilitas

Gambar 3l. Wordcloud Media Massa dari Lainnya

Tabel 4. Daftar Unigram, Bigram, dan Trigram yang dipilih dalam media massa

Kalimat	Kemunculan	Kalimat	Kemunculan	Kalimat	Kemunculan
Unigram		**Bigram**		**Trigram**	
Toyota	4.95%	Honda Brio	0.55%	Toyota Kijang Innova	0.25%
Harga	3.14%	Toyota Agya	0.49%	Simulasi Kredit Cicil	0.17%
Daihatsu	1.79%	Daihatsu Ayla	0.45%	Cek Promo Mei	0.11%
Baru	1.65%	Toyota Avanza	0.41%	Promo Mei Simulasi	0.11%
Honda	1.31%	Toyota Yaris	0.37%	Mei Simulasi Kredit	0.11%
Listrik	0.98%	Toyota Kijang	0.36%	Tepat Oto Oto	0.1%
Mesin	0.94%	Simulasi Kredit	0.34%	Oto Com Banding	0.1%
Bekas	0.91%	Wuling Air	0.33%	Honda Brio Satya	0.08%
Brio	0.85%	Kijang Inova	0.30%	Toyota Astra Motor	0.08%
Carmudi	0.70%	Harga Toyota	0.28%	Harga Bekas Toyota	0.08%

Di media massa, aspek biaya seringkali menghadirkan berita tentang harga kendaraan dan simulasi kredit. Fitur kendaraan tertentu, seperti Wuling Air, serta merek Toyota, menjadi sorotan utama dalam liputan tersebut. Dalam konteks keselamatan, berita seringkali menyoroti merek Daihatsu, Toyota, dan Honda Brio. Isu lingkungan, terutama terkait mobil listrik atau kendaraan listrik, juga sering dibahas, dengan merek Toyota sebagai yang paling banyak muncul. Pada topik reliabilitas, komponen mesin dari merek Toyota mendapat perhatian yang signifikan dalam berbagai liputan.

III.2. Hasil Analisis Model

Rata rata skor *coherence* dari setiap topik pada data media massa dan media sosial diperoleh sebesar 0.37. Secara spesifik, rata rata skor *coherence* yang diperoleh pada data media massa adalah sebesar 0.31, sementara skor coherence yang diperoleh pada data media sosial adalah sebesar 0.34.

Koherensi yang menurun ini membuktikan jika konteks atau topik pembicaraan dalam media sosial dan media massa memiliki hubungan yang cukup erat.

Tabel 5. Hasil Analisis Model LDA data Media Sosial

Nomor Topik	Nama Topik	Inisiasi Kata
1	Biaya	'Pajak', 'Biaya', 'Harga', 'Sparepart', 'Angsuran', 'Kredit', 'Diskon', 'Promo', 'Cicilan', 'Tunai', 'Konsumsi', 'Estimasi', 'Bayar', 'Premi', 'Pemakaian', 'Pengeluaran', 'Pemeliharaan', 'Ritel', 'Penawaran', 'Pembelian', 'Transaksi', 'Inden', 'Subsidi'
2	Fitur	'Kenyamanan', 'Comfort', 'Ruang', 'Desain', 'Aesthetic', 'Audio', 'Sound', 'Speaker', 'Teknologi', 'Infotainment', 'Facelift', 'Spesifikasi', 'Specs', 'Navigasi', 'Sunroof', 'Kursi', 'Layar', 'Monitor', 'Pintu', 'Kaca', 'Kunci', 'Remote', 'Sistem', 'Alarm', 'AC', 'HVAC', 'Sirkulasi', 'Audio', 'MP3', 'Bluetooth', 'Koneksi', 'USB', 'Charger', 'Penyimpanan', 'Boks', 'Kompartemen', 'Kanopi'
3	Keselamatan	'Keamanan', 'Airbag', 'ABS', 'Uji', 'Crash', 'Sabuk', 'Rem', 'Penguncian', 'Tabrakan', 'Pengendalian', 'Kendali', 'Pre-collision', 'Deteksi', 'Perlindungan', 'Impact', 'Struktur', 'Rangka', 'Safety', 'Keselamatan', 'Collision', 'Pengaman'
4	Lingkungan	'Efisiensi', 'Fuel', 'Irit', 'Emisi', 'Eco', 'Green', 'Hybrid', 'Electric', 'EV', 'Hidrogen', 'CO2', 'Gas', 'Biodiesel', 'Karbon', 'Solar', 'Biofuel', 'Propana', 'Clean', 'Ramah', 'Rechargeable', 'Power', 'Air', 'Listrik', 'Biodegradable', 'Eco-friendly'
5	Reliabilitas	'Layanan', 'Customer', 'Kemudahan', 'Servis', 'Maintenance', 'Ban', 'Akselerasi', 'Performa', 'Mesin', 'Kapasitas', 'Penjualan', 'Penjaminan', 'Kualitas', 'Jaminan', 'Kendala', 'Garansi', 'Komponen', 'Durabilitas', 'Tahan', 'Teruji', 'Kepuasan', 'Kredibilitas', 'Kepercayaan', 'Keandalan', 'Perbaikan', 'Purna', 'Perlindungan', 'Pengembalian', 'Kehandalan', 'Reputasi', 'Mesin'
6	Lainnya	'Modifikasi', 'Custom', 'Tuning', 'Upgrade', 'Bekas', 'Second', 'Pre-owned', 'Secondhand', 'Lelang', 'Rumor', 'Tidak Valid', 'Penipuan', 'Kemacetan', 'Klub', 'Komunitas', 'Sahabat', 'Tabrak', 'Kecelakaan'

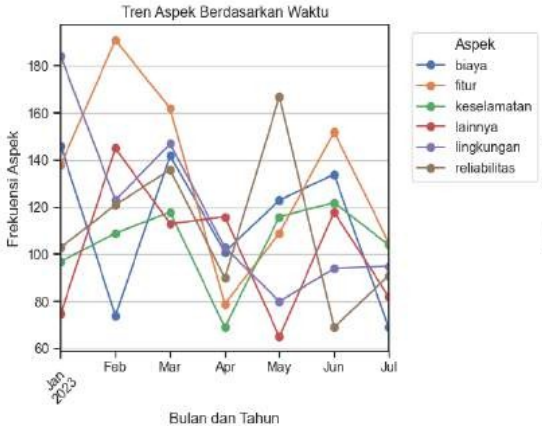
Tabel 6. Hasil Analisis Model LDA

Nomor Topik	Nama Topik	Coherence score		
		Media Massa	Media Sosial	Gabungan
1	Biaya	0.35	0.26	0.38
2	Fitur	0.35	0.37	0.31
3	Keselamatan	0.31	0.26	0.35
4	Lingkungan	0.38	0.27	0.46
5	Reliabilitas	0.46	0.29	0.35
6	Lainnya	0.39	0.33	0.39
	Rata rata	0.34	0.29	0.37

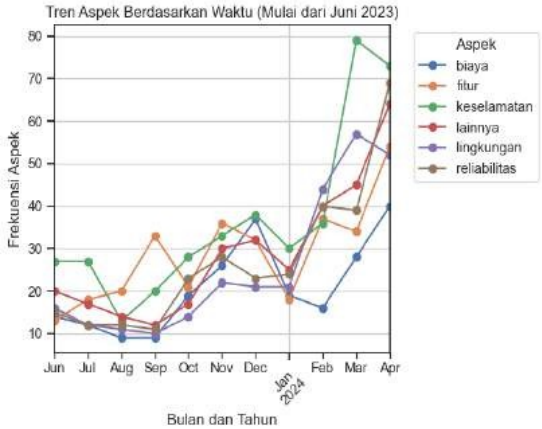
Analisis koherensi topik dari Media Massa, Media Sosial, dan Gabungan menunjukkan bahwa skor koherensi di bawah 0.3 pada Media Sosial, seperti pada topik "Biaya" dan "Keselamatan" (0.26), mengindikasikan tingkat keberagaman data yang tinggi. Sebaliknya, tingkat koherensi yang cukup

tinggi pada Media Massa (rata-rata 0.34) menunjukkan bahwa berita di media massa memiliki tingkat keberagaman yang lebih rendah, dengan narasi yang lebih terstruktur dan konsisten. Gabungan data dari kedua sumber menghasilkan rata-rata koherensi tertinggi (0.37), dengan topik "Lingkungan" paling signifikan (0.46), menunjukkan bahwa integrasi data dari berbagai sumber dapat meningkatkan kualitas analisis topik secara keseluruhan.

III.3. *Trend and Insight*



Gambar 3m. Tren Topik Media Massa



Gambar 3n. Tren Topik Media Sosial

Dari hasil analisis tren yang dilakukan pada data media sosial, terlihat bahwa dari bulan Juni hingga April, terdapat indikasi yang menunjukkan adanya interaksi antar aspek dalam industri otomotif. Temuan ini mengungkapkan bahwa beberapa topik yang dibahas di platform media sosial saling mempengaruhi, menunjukkan korelasi yang signifikan di antara mereka. Di sisi lain, analisis *trend* pada data media massa menunjukkan pola yang berbeda, di mana meskipun beberapa topik mungkin mendapat sorotan individual, tidak terlihat interaksi yang kuat antara topik-topik tersebut.

IV. KESIMPULAN

Dalam penelitian, telah dijelaskan secara garis besar klasifikasi topik, dan *trend*-nya dalam industri otomotif. Telah diaplikasikan *semi-supervised* LDA sebagai algoritma untuk melakukan topic modelling, dengan menggunakan data 28.934 data yang bersumber dari 15 situs web media massa, dan 6 *platform* media sosial yang nantinya dilakukan klasifikasi menjadi 6 topik, yaitu: Biaya, Reliabilitas, Keamanan, Keselamatan, Fitur, dan Lainnya. Penelitian yang telah dilakukan ini juga dapat memberikan manfaat bagi para industri otomotif, seperti Toyota, Hyundai, dan lain lain, karena dapat membantu untuk mengklasifikasikan topik dari data yang tidak teratur dan beragam.

Pada analisis yang telah dilakukan, dapat kita ambil kesimpulan jika antar aspek pada media sosial memiliki interaksi yang signifikan. Sebaliknya, media massa cenderung lebih fokus untuk melakukan penyebaran informasi terkait promosi dan harga kendaraan, serta kemajuan teknologi kendaraan listrik. Dari temuan ini dapat disarankan bagi agar *brand* otomotif dapat menyempurnakan strategi dalam pembuatan konten yang relevan di media sosial, karena dengan konten yang relevan dan mengikuti topik trend yang sedang terjadi, konten yang dihasil dapat menarik lebih banyak perhatian, sehingga meningkatkan keterlibatan pengguna.

Coherence score rata-rata untuk topik Gabungan (0.37) lebih tinggi dibandingkan dengan Media Massa (0.34) dan Media Sosial (0.29), menunjukkan bahwa integrasi data dari kedua sumber

memberikan pemahaman yang lebih baik tentang topik. Topik-topik yang sering dibahas mencakup Biaya, Reliabilitas, Keselamatan, dan Lingkungan, yang merupakan area utama perhatian konsumen. Hal ini memberikan peluang bagi perusahaan otomotif untuk mengembangkan strategi pemasaran dan pengembangan produk yang lebih tepat sasaran, berdasarkan tren dan perhatian konsumen yang teridentifikasi melalui analisis data ini.

Dari penelitian ini ditunjukan bahwa merek-merek otomotif seperti toyota, honda dan daihatsu sering menjadi topik perbincangan dan informasi di kedua jenis media. topik pada yang paling banyak muncul berkaitan dengan harga, fitur, kendaraan, reliabilitas mesin. Maka disarankan bagi brand lain dalam industri otomotif untuk fokus menciptakan suatu pembeda yang unik, dengan menawarkan solusi yang lebih inovatif. Menonjolkan fitur dan keunggulan yang tidak dimiliki oleh merek yang mendominasi, dapat membuat konsumen menjadi lebih tertarik akan pilihan baru yang telah ditawarkan.

REFERENSI

1. L. Sun and Y. Yin, "Discovering themes and trends in transportation research using topic modeling," *Transp. Res. Part C Emerg. Technol.*, vol. 77, pp. 49–66, 2017, doi: 10.1016/j.trc.2017.01.013.
2. M. McCombs, "The agenda-setting role of the mass media in the shaping of public opinion". In: Mass Media Economics Conference. London School of Economics, London," *Infoamerica*, no. January 2011, p. 22, 2002.
3. B. Kim, A. Xiong, D. Lee, and K. Han, "A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions," *PLoS One*, vol. 16, no. 12 December, pp. 1–28, 2021, doi: 10.1371/journal.pone.0260080.
4. R. K. Nielsen and L. Graves, "Audience perspectives on fake news," *Reuters Inst. Study Journal.*, no. October, pp. 1–8, 2017.
5. H. Allcott and M. Gentzkow, "Nber Working Paper Series Social Media and Fake News in the 2016 Election," *J. Econ. Perspect.*, vol. 31, no. 2, pp. 211–236, 2017, [Online]. Available: https://web.stanford.edu/~gentzkow/research/fakenews.pdf
6. B. Yang, R. Zhang, X. Cheng, and C. Zhao, "Exploring information dissemination effect on social media: an empirical investigation," *Pers. Ubiquitous Comput.*, vol. 27, no. 4, pp. 1469–1482, 2023, doi: 10.1007/s00779-023-01710-7.
7. M. Liebenlito, N. Inayah, E. Choerunnisa, T. E. Sutanto, and S. Inna, "Active Learning on Indonesian Twitter Sentiment Analysis Using Uncertainty Sampling," *J. Appl. Data Sci.*, vol. 5, no. 1, pp. 114–121, 2024, doi: 10.47738/jads.v5i1.144.
8. D. Wang, M. Thint, and A. Al-Rubaie, "Semi-supervised latent Dirichlet allocation and its application for document classification," *Proc. 2012 IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol. Work. WI-IAT 2012*, pp. 306–310, 2012, doi: 10.1109/WI-IAT.2012.211.
9. H. Soleimani and D. J. Miller, "Semi-supervised multi-label topic models for document classification and sentence labeling," *Int. Conf. Inf. Knowl. Manag. Proc.*, vol. 24-28-Octo, pp. 105–114, 2016, doi: 10.1145/2983323.2983752.
10. Y. Zhang and W. Wei, "A jointly distributed semi-supervised topic model," *Neurocomputing*, vol. 134, pp. 38–45, 2014, doi: 10.1016/j.neucom.2012.12.077.
11. W. Xu, H. Sun, C. Deng, and Y. Tan, "Variational autoencoder for semi-supervised text classification," *31st AAAI Conf. Artif. Intell. AAAI 2017*, pp. 3358–3364, 2017, doi: 10.1609/aaai.v31i1.10966.
12. C. Zhou, H. Ban, J. Zhang, Q. Li, and Y. Zhang, "Gaussian Mixture Variational Autoencoder for Semi-Supervised Topic Modeling," *IEEE Access*, vol. 8, pp. 106843–106854, 2020, doi: 10.1109/ACCESS.2020.3001184.