Terakreditasi SINTA Peringkat 3
Surat Keputusan Direktur Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Nomor: 72/E/KPT/2024
masa berlaku mulai Volume 10 Nomor 1 Tahun 2023 sampai Volume 14 Nomor 2 Tahun 2027

Terbit online pada laman web jurnal: **https://jurnal.plb.ac.id/index.php/tematik/index**

TEMATIK

Jurnal **Teknologi Informasi Komunikasi (e-Journal)**

Vol. 11 No. 2 (2024) 195 - 203 ISSN Media Elektronik: 2443-3640

Identifikasi Opini Publik Terhadap Kendaraan Listrik dari Data Komentar YouTube: Pemodelan Topik Menggunakan BERTopic

Identifying Public Opinion on Electric Vehicles from YouTube Comment Data: Topic Modelling Using BERTopic

Kristine Angelina Simanjuntak[1], Muhamad Koyimatu[2], Yolla Putri Ervanisari[3], Tasmi[4]
[1,2,3,4]Ilmu Komputer, Sains dan Ilmu Komputer, Universitas Pertamina
[1]kristineangelina12@gmail.com, [2]koyimatu@universitaspertamina.ac.id, [3]yolaep.yp@gmail.com

Abstract

The Indonesian government is encouraging the transition to electric vehicles to reduce the use of fossil fuels and the negative environmental impact. This transition sparked controversy because Indonesia is still heavily dependent on coal-fired power plants, and many argue that the transition is not ready without adequate renewable energy and supporting infrastructure. Public opinion analysis is crucial in considering the introduction of electric vehicles in Indonesia due to the controversial nature of the transition. The opinion is transmitted through YouTube by taking comment data, then grouped into a topic to identify public opinion. The topic modeling method used is a BERTopic transformer model using IndoBERTweet in embedding. Once public opinion is modeled into a topic, changes in public opinion are evaluated using coherence score metrics and topic diversity as a measure of the consistency and diversity of the topic. The resulting topics have a coherence value of around 0.6 to 1 and a diversity value of 0.95838. This indicates that the resulting themes have strong semantic similarities and high diversity in terms of word usage and capture various aspects of text documents well.

Keywords: BERTopic, coherence score, electric vehicles, public opinion, topic modeling

Abstrak

Pemerintah Indonesia mendorong transisi ke kendaraan listrik untuk mengurangi penggunaan bahan bakar fosil dan dampak negatif lingkungan. Namun, transisi ini memicu perdebatan karena Indonesia masih sangat bergantung pada pembangkit listrik berbahan bakar batu bara, dan banyak yang berpendapat bahwa transisi ini belum siap tanpa energi terbarukan yang memadai dan infrastruktur pendukung. Aspek sosial dan penerimaan masyarakat menjadi kunci dalam mempercepat adopsi kendaraan listrik. Analisis opini publik sangat penting dalam mempertimbangkan pengenalan kendaraan listrik di Indonesia karena sifat kontroversial dari transisi. Opini tersebut disampaikan melalui *YouTube* dengan mengambil data komentar, kemudian dikelompokkan menjadi sebuah topik untuk mengidentifikasi opini masyarakat. Metode *topic modeling* yang dipergunakan adalah model *transformer BERTopic* dengan memanfaatkan *IndoBERTweet* dalam melakukan *embedding.* Setelah opini publik dimodelkan menjadi sebuah topik, perubahan opini publik dievaluasi menggunakan metrik *coherence score* dan *topic diversity* sebagai pengukur koherensi dan keberagaman dari topik. Topik-topik yang dihasilkan memiliki nilai koherensi disekitar angka 0.6 hingga 1 dan nilai *diversity* sebesar 0.95838. Hal tersebut menunjukkan topik yang dihasilkan memiliki kesamaan semantik yang kuat dan keragaman yang tinggi dalam hal penggunaan kata-kata dan menangkap berbagai aspek dari dokumen teks dengan baik

Kata kunci: *BERTopic, coherence score,* kendaraan listrik, opini publik, *topic modeling*

1. Pendahuluan

Pemerintah Indonesia sedang melakukan transisi ke kendaraan listrik untuk mengurangi penggunaan kendaraan berbahan bakar fosil, karena bahan bakar fosil memberikan dampak negatif terhadap lingkungan. Salah satu langkah yang diambil adalah dengan menetapkan Peraturan Presiden Nomor 55 Tahun 2019 tentang Percepatan Program Kendaraan Bermotor Listrik Berbasis Baterai (KBLBB) [1]. Namun, transisi ini memicu perdebatan mengenai dampak ekonomi dan lingkungan, karena sebagian besar sumber listrik masih berasal dari batu bara, sedangkan batu bara dapat meningkatkan emisi CO_2 [2].

Diterima Redaksi: 12-11-2024 | Selesai Revisi: 12-11-2024 | Diterbitkan Online: 20-12-2024

Indonesia menghadapi tantangan besar dalam transisi ke kendaraan listrik, seperti kebutuhan energi yang masih bergantung pada batu bara, infrastruktur pengisian daya yang terbatas, harga tinggi, dan kurangnya fasilitas pendukung seperti bengkel listrik [3]. Masyarakat juga memiliki pandangan yang beragam; meskipun ada yang melihat kendaraan listrik sebagai solusi untuk mengurangi ketergantungan pada minyak, ada juga yang khawatir tentang kemacetan dan kesiapan infrastruktur [4].

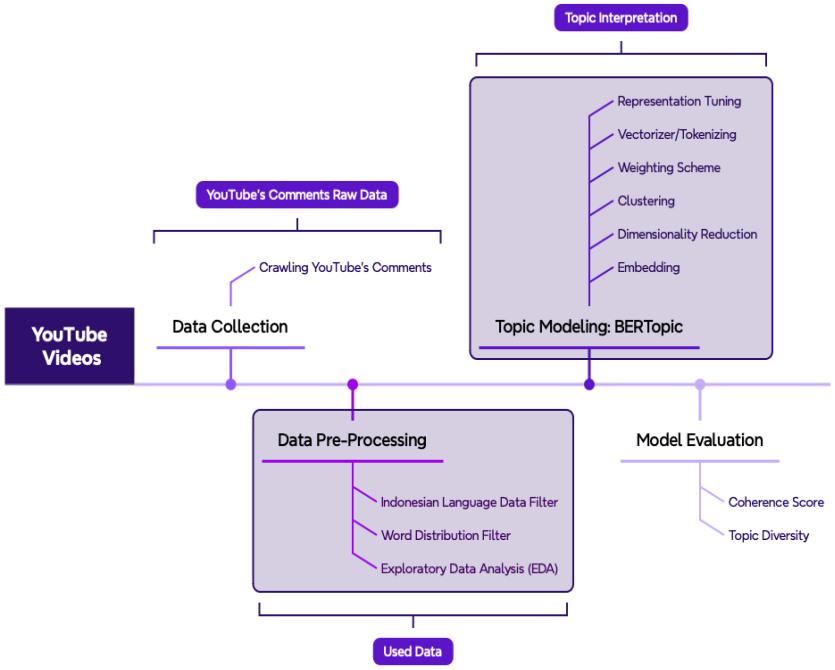
Dengan adanya beragam pandangan masyarakat terkait kendaraan listrik tersebut, maka pada penelitian ini bertujuan untuk memahami opini publik mengenai kendaraan listrik di Indonesia melalui analisis komentar pada video *YouTube*. Opini masyarakat dapat memberikan wawasan penting bagi pembuat kebijakan untuk mengatasi hambatan dan mempromosikan penerimaan kendaraan listrik [3][5].

Terdapat peneliti yang telah mengkaji topik terkait kendaraan listrik yaitu Suresha (2021) melakukan pemodelan topik terhadap kendaraan listrik dengan menerapkan teknik pemodelan *Latent Dirichlet Allocation (LDA)* dengan menggunakan data *Twitter*, dalam tulisannya Suresha menjelaskan *LDA* kesulitan dalam menginterpretasi topik dan terbatas dalam mempertimbangkan urutan kata, karena *LDA* menganggap setiap kata dalam dokumen sebagai *independent* satu sama lain, sehingga hal tersebut mempengaruhi akurasi identifikasi topik. Pada tahun 2023, Ogunleye melakukan penelitian dan hasilnya menunjukkan bahwa *KernelPCA* dan *K-means* dalam arsitektur *BERTopic* menghasilkan topik yang koheren dengan skor koherensi 0.8463, serta menangani batasan *LDA* dalam interpretasi topik dan pertimbangan urutan kata [6].

Oleh karena itu, dalam penelitian ini memanfaatkan *BERTopic* sebagai teknik pemodelan topik yang mampu mengatasi permasalahan *LDA*, sesuai untuk menganalisis data tentang kendaraan listrik di Indonesia agar mendapatkan pemahaman yang lebih mendalam tentang isu-isu utama yang dihadapi, mengidentifikasi kebutuhan masyarakat yang belum terpenuhi, mengembangkan strategi yang lebih efektif dalam meningkatkan adopsi kendaraan listrik, dan mengatasi hambatan-hambatan yang ada. Data-data ini dapat memberikan wawasan kepada pemerintah, produsen mobil listrik, dan masyarakat umum untuk mengambil langkah-langkah yang tepat dalam menghadapi tantangan dan kesempatan yang terkait dengan perkembangan kendaraan listrik.

2. Metode Penelitian

Gambar 1 menjelaskan mengenai alur metodologi penelitian yang dilakukan dalam penelitian.



Gambar 1. *Flowchart* Metodologi Penelitian

2.1. *Data Colllection*

Penelitian ini merupakan sebuah modifikasi atau lanjutan dalam penulisan yang dilakukan oleh [7]. Pengumpulan data dilakukan menggunakan metode *web crawling.* Data yang dikumpulkan berupa komen publik dari berbagai video *youtube* yang muncul sesuai dengan kata kunci yang diberikan. Tabel 1 menunjukkan pengumpulan data video mengambil dari tahun 2018 hingga tahun 2024.

DOI: https://doi.org/10.38204/tematik.v11i2.2096

Pengumpulan data komentar menggunakan *YouTube Data API* yang disediakan oleh *Google. YouTube Data API* mengizinkan pengguna untuk mengakses komentar-komentar dengan memasukkan *video id* yang ada di setiap video *YouTube*, kemudian *API* mengumpulkan data komentar tersebut kedalam sebuah *file CSV* terlihat pada Tabel 1.

Tabel 1. Rincian Jumlah Data Video dan Komentar Setiap Tahun

Tahun	Jumlah Video	Jumlah Data Komentar
2018	548	851
2019	96	145
2020	144	218
2021	7770	10392
2022	3508	4719
2023	5943	9038
2024	4571	5918

2.2. *Data Pre-Processing*

Setelah proses *crawling* telah dilakukan, *raw dataset* atau kumpulan data mentah yang tidak memiliki format yang teratur dibersihkan terlebih dahulu dengan tahap *pre-processing. Pre-processing* dilakukan agar meningkatkan kualitas data dengan mempersiapkan *raw dataset* agar dapat diolah lebih lanjut untuk menghasilkan analisis yang lebih akurat dan bermakna. Beberapa tahapan *pre-processing* yang dilakukan antara lain:

Indonesian Language Data Filter: Melakukan proses menseleksi data komentar pada *DataFrame* yang hanya berbahasa Indonesia menggunakan *library langid.*

Word Distribution Filter: Melakukan proses menseleksi data komentar yang hanya mengandung minimal delapan kata menggunakan *word_distribution* yang berfungsi untuk menghitung distribusi kata dalam sebuah kalimat.

Exploratory Data Analysis (EDA): Melakukan proses awal untuk menganalisis data untuk memahami struktur dan isi dari data. *EDA* penting untuk mempersiapkan data sebelum melanjutkan ke tahap analisis lebih lanjut atau ke dalam tahap pemodelan. Dengan menggunakan *plot* dan grafik untuk memahami distribusi, struktur pola, dan hubungan antar variabel dalam data.

2.3. *Topic Modeling: BERTopic*

Tahapan-tahapan *topic modeling* yang digunakan dalam penelitian terdiri dari proses-proses yang dijelaskan oleh dokumentasi *github* yang ditulis oleh Maarten Grootendorst dalam *paper* yang berjudul "BERTopic: Neural topic modeling with a class-based TF-IDF procedure" [8].

BERTopic bekerja dengan mengubah dokumen menjadi nilai numerik, yang disebut *embeddings*. Representasi numerik dilakukan agar dapat diolah oleh algoritma pengelompokkan dan pemodelan topik. Proses ini mengubah kalimat menjadi kumpulan vector yang dapat digunakan untuk mengindentifikasi semantik dan kemiripan antar kalimat. Model *embedding* yang digunakan adalah *IndoBERT* sebagai model *embedding* Bahasa Indonesia yang dibangun menggunakan metode *transfer learning* dari model *BERT*.

Dalam *BERTopic*, algoritma *dimensionality reduction* digunakan untuk mengurangi jumlah dimensi atau fitur dalam sekumpulan data untuk mencegah masalah yang muncul ketika bekerja dengan ruang yang berdimensi tinggi. Proses *dimensionality reduction* dilakukan dengan tujuan untuk mengurangi kompleksitas data dan menghapus data yang tidak relevan agar lebih mudah untuk melakukan visualisasi pada data.

Clustering dilakukan untuk membantu memahami dan mengidentifikasi topik-topik yang terdapat dalam data dengan mengelompokkan data dengan sifat yang sama ke dalam kelompok-kelompok kecil. *Clustering* sebagai pengontrol jumlah topik dengan menggunakan parameter *n_clusters.* Hal tersebut merupakan parameter yang mendukung pembuatan jumlah topik yang tetap.

Dengan menggunakan *CountVectorizer*, dapat dilakukan beberapa hal seperti: Menghapus *stopwords* kata-kata umum yang tidak memberi makna seperti "dan", "atau", dan sebagainya; Mengabaikan kata-kata yang jarang muncul dan tidak relevan; Tokenisasi, memecah teks menjadi kata-kata atau token dengan memisahkan teks menjadi unit-unit yang lebih kecil seperti berdasarkan spasi atau pola tertentu.

Weighting scheme adalah teknik yang digunakan untuk mengurangi pengaruh kata yang tidak relevan dan mengemukakan kata yang lebih relevan dalam proses pemodelan topik. Dalam *BERTopic*, *weighting scheme* terdiri atas dua bagian: *term frequency (tf)* dan *inverse document frequency (idf)*. *Term frequency (tf)* adalah frekuensi kata dalam *cluster*, yang menggambarkan berapa banyak kali kata tersebut muncul dalam *cluster* tersebut. *Inverse document frequency (idf)* adalah logaritma dari 1 *plus* jumlah *cluster* yang mengandung kata tersebut, dibagi dengan jumlah *cluster* yang mengandung kata tersebut. *Representation default* dari topik dihitung melalui *c-TF-IDF* namun, *c-TF-IDF* diperkuat oleh *CountVectorizer* dengan mengubah teks menjadi *representasi bag-of-words*, dilakukan dengan menghitung frekuensi kata-kata.

Proses mengubah representasi topik yang dihasilkan oleh model *BERTopic*. *Representation tuning* dapat dilakukan dengan menggunakan beberapa model yang telah terimplementasikan dalam *BERTopic*, seperti *MaximalMarginalRelevance, OpenAI, KeyBERTInspired*, dan lain-lain. *OpenAI* adalah model yang menggunakan *API OpenAI* untuk mengelompokkan topik untuk mengelompokkan topik yang lebih baik dan memperbaiki kualitas topik.

DOI: https://doi.org/10.38204/tematik.v11i2.2096

KeyBERTInspired adalah model yang menggunakan algoritma *KeyBERT* untuk membantu memperbaiki kualitas topik dan memperingkatkan koherensi topik.

2.4. *Model Evaluation*

Hasil dari pemodelan topik dapat dievaluasi dan dibandingkan dengan menggunakan beberapa metrik evaluasi. Metrik seperti *coherence score* dan *topic diversity* banyak diterapkan untuk digunakan sebagai metrik evaluasi [9].

Coherence score dalam *BERTopic* adalah metrik yang digunakan untuk mengukur kemiripan antara kumpulan kata dalam sebuah topik dengan kumpulan kata yang dapat dijumpai dalam dokumen lain. *Coherence score* yang dihasilkan berupa skala dari 0 hingga 1 di mana konsisten yang baik (kesamaan tinggi) memiliki skor dari 1, dan konsistenitas yang buruk (kesaman rendah) mempunyai skor dari 0. Metrik evaluasi menggunakan *coherence score* menggunakan rumus *UMass Coherence* dalam Persamaan 1 untuk mengukur kualitas topik.

$$C_{UMass} = \sum_{m=2}^{M}\sum_{l=1}^{m-1} log\frac{D(w_m,w_l)+\epsilon}{D(w_l)} \quad (1)$$

w_m, w_l adalah kata-kata dalam topik, $D(w_m, w_l)$ adalah Jumlah dokumen yang mengandung kedua kata w_m *dan* w_l, $D(w_l)$ adalah Jumlah dokumen yang mengandung kata w_l, ϵ adalah *Smoothing* parameter untuk menghindari log(0), M adalah Jumlah kata dalam topik.

Untuk menghitung *UMass coherence score*, langkah pertama adalah mengumpulkan frekuensi kata dari dokumen, menghitung jumlah dokumen yang mengandung setiap kata dan pasangan kata. Selanjutnya, hitung frekuensi kemunculan bersama setiap pasangan kata dalam topik. Untuk menghindari logaritma dari nol, tambahkan parameter *smoothing* ϵ. Kemudian, hitung logaritma dari rasio antara *co-occurrence* dan frekuensi kata individu. Terakhir, jumlahkan semua nilai logaritma rasio untuk semua pasangan kata dalam topik. Hasilnya adalah *UMass coherence score* yang menunjukkan koherensi topik.

Topic diversity dikaitkan dengan eksekusi spekulatif yang dapat dikendalikan oleh pengguna, di mana keragaman dalam *cluster* digunakan untuk membedakan *nodes* dengan potensi untuk perbaikan dari *nodes* topik spesifik [10]. Nilai dari *topic diversity* berkisar dari 0 hingga 1, dengan nilai yang lebih rendah atau mendekati 0 menunjukkan topik yang berlebihan dan nilai yang tinggi atau mendekati 1 menunjukkan keragaman topik yang lebih baik [8]. Untuk menghitung metrik evaluasi *topic diversity*, rumus yang digunakan adalah Persamaan 2.

$$Topic\ Diversity = 1 - \frac{1}{\mathrm{N}}\sum_{i=1}^{N}\frac{|W_i|}{W} \quad (2)$$

N adalah Jumlah topik, $|W_i|$ adalah Jumlah kata unik dalam topik I, W adalah Total jumlah kata unik di semua topik.

Untuk menghitung *topic diversity*, langkah pertama adalah mengidentifikasi kata-kata unik dalam setiap topik dan menghitung jumlahnya. Kemudian, hitung total kata unik dari semua topik. Selanjutnya, bagi jumlah kata unik dalam setiap topik dengan total kata unik untuk mendapatkan proporsi kata unik per topik. Setelah itu, rata-rata proporsi ini dibagi dengan jumlah topik untuk mendapatkan nilai rata-rata normalisasi. Akhirnya, kurangi nilai rata-rata normalisasi dari satu untuk mendapatkan nilai *topic diversity*. Nilai ini menunjukkan seberapa beragam topik yang dihasilkan oleh model.

3. Hasil dan Pembahasan

Rangkaian

3.1. *Data Collection*

Proses pengumpulan data dengan metode *crawling* menggunakan sumber data penelitian dari media sosial *YouTube.* Pengumpulan data komentar masyarakat menggunakan *keyword* 'Kendaraan Listrik di Indonesia' dan di 'sort by relevance' *filter* untuk hanya mengambil video yang relevan dengan *keyword* saja. Video-video yang dipilih hanya berbahasa Indonesia, memiliki kualitas audio-visual yang baik, dan bersifat edukatif atau informatif. Isi konten dari video yang diambil harus mengandung tiga kriteria antara lain: Menjelaskan tentang adopsi kendaraan listrik di Indonesia oleh berbagai perusahaan mobil luar negeri ataupun dalam negeri; Membahas kebijakan pemerintah Indonesia terkait kendaraan listrik; Membahas perkembangan kendaraan listrik di Indonesia.

Setelah pencarian video yang memenuhi ketiga kriteria diatas dilakukan, maka diperoleh 20 video. Proses *crawling* yang dilakukan dimulai pada tanggal 1 Juni 2024.

	author	text
0	@lazyside5452	Indonesia bukan kekurangan orang cerdas, Indon...
1	@imambbmsarjanapedangdut7886	PEMERINTAH INDONESIA HARUS DISUAP BARU APAPUN ...
2	@user-bv7tk4sn3x	Mobil yang terbakar non listrik seperti kasus ...
3	@taslimhobibonsai3821	Ooh..., paok... paok... 😇
4	@siiikicen6320	semoga di kepresidenan pak prabowo gibran bisa...
...	...	...
36891	@jayjolupoi88891	kasihan, sulawesi, dikeruk habis habisan tapi ...
36892	@Surya-kh3kc	Jokowi emank presiden sesat\nBikin cina kaya t...
36893	@wulanalgafari8846	Di Sulawesi, di Halmahera semua dikeruk habis2an.
36894	@LumburAI	Narasi kok jadi gini \nBerasa TV one 5 tahun l...
36895	@nigellaid	Apakah ini kendaraan Listrik selalu dipromosik...

36896 rows × 7 columns

Gambar 2. Hasil *Data Collection*

Gambar 2 menunjukkan hasil *crawling*, sehingga data yang diperoleh sebanyak 36.896 data komentar, hasil *crawling* tersebut merupakan sebuah data mentah yang

DOI: https://doi.org/10.38204/tematik.v11i2.2096

nantinya di *pre-processing* (tahapan pembersihan data) sebelum digunakan ke tahapan pemodelan.

3.2. *Data Pre-Processing*

Sesudah proses *crawling* data berhasil, kemudian hasil *crawling* data tersebut disimpan ke dalam sebuah *data frame* untuk memudahkan proses pembersihan data atau proses *preprocessing* dilakukan. Beberapa tahapan *pre-processing* yang dilakukan antara lain:

Indonesian Language Data Filter: Tahapan pertama dalam *pre-processing*, *data frame* yang masih tidak terstruktur dari hasil *crawling* di *filter* hanya komentar yang menggunakan bahasa Indonesia dengan hasil akhir ditunjukkan oleh Gambar 3 dengan jumlah data yang diperoleh sebanyak 31.281 data, dengan kata lain ada sebanyak 5.615 data yang tidak berbahasa Indonesia.

	author	text
0	@lazyside5452	Indonesia bukan kekurangan orang cerdas, Indon...
2	@user-bv7tk4sn3x	Mobil yang terbakar non listrik seperti kasus ...
4	@siiikicen6320	semoga di kepresidenan pak prabowo gibran bisa...
5	@GeminiPonsel-ym1rc	yah gmna ya...klk barang dari luwar enak pajak...
6	@ryoarcho9089	Langsung aja ya ke intinya.. \n\n"PEMERINTAH K...
...	...	...
36891	@jayjolupoi88891	kasihan, sulawesi, dikeruk habis habisan tapi ...
36892	@Surya-kh3kc	Jokowi emank presiden sesat\nBikin cina kaya t...
36893	@wulanalgafari8846	Di Sulawesi, di Halmahera semua dikeruk habis2an.
36894	@LumburAI	Narasi kok jadi gini \nBerasa TV one 5 tahun l...
36895	@nigellaid	Apakah ini kendaraan Listrik selalu dipromosik...

31281 rows × 7 columns

Gambar 3. Hasil *Data Pre-Processing* Menggunakan *Indonesian Language Data Filter*

Word Distribution Filter: Tahapan kedua dalam *pre-processing* dilakukan proses menyeleksi data komentar berdasarkan jumlah kata yang dimiliki. Pemilihan data komentar yang ingin diseleksi pada setiap komen didasarkan pada perhitungan *word distribution* untuk melihat distribusi kata dalam setiap kalimat data komentar. Hal tersebut diperlihatkan pada Gambar 4, yang menunjukkan bahwa pada kuartil pertama dari data (25% bagian dari total keseluruhan data komentar) memiliki data komentar yang mengandung delapan kata, sehingga data komentar yang diambil hanya data komentar yang mengandung minimal delapan kata dalam suatu kalimat.

	like_count	word_distribution
count	31247.000000	31247.000000
mean	1.967901	20.388773
std	30.585866	25.042397
min	0.000000	1.000000
25%	0.000000	8.000000
50%	0.000000	13.000000
75%	0.000000	24.000000
max	2646.000000	654.000000

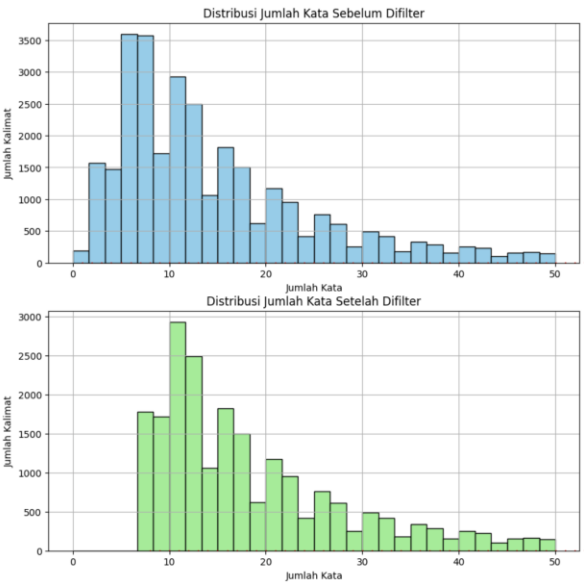
Gambar 4. Distribusi Kata dalam Data Komentar

Proses *word distribution filter* ditunjukkan pada Gambar 5 dengan jumlah data yang diperoleh sebanyak 22.649 data yang berarti ada sebanyak 8.632 data yang tidak memenuhi syarat penyeleksian jumlah kata.

	author	text
0	@lazyside5452	Indonesia bukan kekurangan orang cerdas, Indon...
1	@user-bv7tk4sn3x	Mobil yang terbakar non listrik seperti kasus ...
2	@siiikicen6320	semoga di kepresidenan pak prabowo gibran bisa...
3	@GeminiPonsel-ym1rc	yah gmna ya...klk barang dari luwar enak pajak...
4	@ryoarcho9089	Langsung aja ya ke intinya.. \n\n"PEMERINTAH K...
...	...	...
31275	@RudiDwiHartanto	Penting rasanya melihat Nickel boom dengan per...
31276	@jayjolupoi88891	kasihan, sulawesi, dikeruk habis habisan tapi ...
31277	@Surya-kh3kc	Jokowi emank presiden sesat\nBikin cina kaya t...
31279	@LumburAI	Narasi kok jadi gini \nBerasa TV one 5 tahun l...
31280	@nigellaid	Apakah ini kendaraan Listrik selalu dipromosik...

22649 rows × 7 columns

Gambar 5. Hasil *Data Pre-Processing* Menggunakan *Word Distribution Filter*



Gambar 6. Grafik Perbandingan Perubahan Jumlah Kata Pada Data Komentar

Penggunaan *word distribution filter* mengurangi *noise*, meningkatkan kualitas topik, dan distribusi kata yang lebih baik, sehingga model dapat fokus pada konten yang lebih informatif. Proses *word distribution filter*

DOI: https://doi.org/10.38204/tematik.v11i2.2096

juga membantu dalam mendapatkan distribusi kata yang lebih kaya dan bervariasi, sehingga mendasar untuk mengidentifikasi topik yang lebih akurat dan bermakna [11]. Gambar 6 menunjukkan perubahan distribusi jumlah kata dalam data komentar sebelum dan setelah dilakukannya *word distribution filter*. Grafik "Distribusi Jumlah Kata Setelah Difilter*"* menunjukkan bahwa tidak ada lagi kalimat yang memiliki jumlah kata dibawah delapan.

Pasca kedua tahapan *data pre-processing* sebelumnya dilakukan, tahap akhir dalam mempersiapkan data bersih yang digunakan sebagai data akhir dalam pemodelan *BERTopic* merupakan penghapusan data duplikat dan data entri yang kosong. Gambar 7 menunjukkan hasil akhir jumlah data yang diperoleh sebanyak 22.580 data komentar.

	author	text
0	@lazyside5452	Indonesia bukan kekurangan orang cerdas, Indon...
1	@user-bv7tk4sn3x	Mobil yang terbakar non listrik seperti kasus ...
2	@siiikicen6320	semoga di kepresidenan pak prabowo gibran bisa...
3	@GeminiPonsel-ym1rc	yah gmna ya...klk barang dari luwar enak pajak...
4	@ryoarcho9089	Langsung aja ya ke intinya.. \n\n"PEMERINTAH K...
...	...	...
22644	@RudiDwiHartanto	Penting rasanya melihat Nickel boom dengan per...
22645	@jayjolupoi88891	kasihan, sulawesi, dikeruk habis habisan tapi ...
22646	@Surya-kh3kc	Jokowi emank presiden sesat\nBikin cina kaya t...
22647	@LumburAI	Narasi kok jadi gini \nBerasa TV one 5 tahun l...
22648	@nigellaid	Apakah ini kendaraan Listrik selalu dipromosik...

22580 rows × 8 columns

Gambar 7. Data Bersih dari Hasil Data *Pre-Processing*

Gambar 8. *Word Cloud* dari Hasil *Data Pre-Processing*

Exploratory Data Analysis (EDA): Gambar 8 menampilkan berbagai kata dalam bahasa Indonesia yang sering digunakan dalam data komentar. Kata-kata yang paling menonjol termasuk "dan," "yang," "untuk," "mobil," dan "listrik," menunjukkan fokus utama pada topik mobil listrik. Selain itu, kata-kata seperti "Indonesia," "bisa," "anak," dan "bangsa" menunjukkan adanya kebanggaan nasional dan harapan bahwa teknologi ini bisa dikembangkan dan dimanfaatkan oleh anak bangsa. Kata "pemerintah" dan "harga" juga sering muncul, mengindikasikan adanya diskusi tentang peran pemerintah dan aspek biaya dalam pengembangan dan adopsi mobil listrik. Kata-kata lain seperti "karena," "kalau," dan "saya" menunjukkan pendapat pribadi dan alasan di balik dukungan atau keraguan terhadap mobil listrik. Secara keseluruhan, *word cloud* ini memberikan gambaran tentang berbagai aspek yang dibicarakan dalam konteks mobil listrik di Indonesia, mencakup aspek teknologi, ekonomi, kebijakan, dan kebanggaan nasional.

3.3. *Embeddings*

IndoBERTweet merupakan sebuah *transformer encoder* dengan 12 *hidden layers* (768 dimensi), 12 *attention heads*, dan 3 *feed-forward hidden layers* (3.072 dimensi). *IndoBERTweet* berguna untuk *embedding* dikarenakan adaptasi model yang efisien di bawah kelemahan kata. Pada penelitian Koto melakukan *train* terhadap lima *IndoBERTweet* model terhadap tujuh *datasets* berbasis *Twitter* menghasilkan *embedding* menggunakan *IndoBERTweet* membuat *pretraining* lima kali lebih cepat dan lebih efektif daripada metode lainnya [12].

Tahapan awal dari pemodelan topik menggunakan *BERTopic* merupakan proses *embeddings*. Proses *embeddings* mengolah data komentar yang telah bersih di *embedded* menggunakan sebuah *pre-trained language model* dalam sebuah kerangka *SentenceTransformer* dengan menggunakan model "indolem/indobertweet-base-uncased". Model ini telah dilatih menggunakan data bahasa Indonesia, termasuk teks-teks dari *Twitter* atau yang sekarang dikenal sebagai *X*. Model ini memiliki pemahaman yang lebih baik tentang struktur bahasa, kata, dan konteks yang sering digunakan dalam percakapan di *platform* media sosial.

Proses *embeddings* dengan menggunakan *IndoBERTweet* menghasilkan vektor 768 dimensi dengan setiap baris dan kolom merepresentasikan fitur atau karakteristik dari satu teks data komentar. Angka desimal yang dihasilkan bernilai positif dan negatif menunjukkan kontribusi masing-masing fitur terhadap representasi keseluruhan dari teks. Gambar 9 menunjukkan hasil dari perubahan data teks (data bersih) menjadi sebuah representasi numerik untuk setiap data komentar.

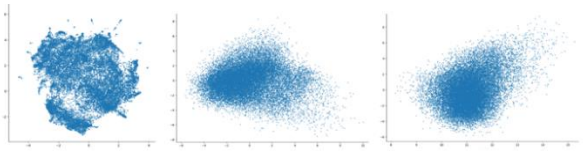
```
          0         1         2         3         4         5         6    \
0  0.311698  0.429043 -0.401786  0.278630 -1.300210 -0.465517  0.879253
1  0.419530  0.068921 -0.836793  0.027632 -0.962753 -0.177591  0.193193
2  0.171250  0.006263 -0.408623  0.072389 -1.166108 -0.689838  0.534683
3  0.495835 -0.314806 -0.498253  0.240747 -0.546063 -0.238465  0.282822
4  0.624659 -0.457541 -0.296998  0.532048 -0.629165 -0.246349 -0.215029

          7         8         9    ...       758       759       760       761  \
0  0.135433 -0.635349  0.347769  ... -0.078027 -0.178058  1.375720  0.267495
1  0.911291 -0.343007  0.521541  ... -0.220585 -0.066101  0.987454  0.106156
2 -0.097023 -0.208324 -0.054120  ... -0.325402 -0.343624  0.729028 -0.614732
3  0.674118  0.071402  0.484644  ...  0.124661 -0.181628  0.545955 -0.332737
4  0.666674  0.175833  0.162892  ...  0.489318 -0.369841  0.822919  0.071634

          762       763       764       765       766       767
0 -0.751608  0.204557 -0.823824  0.073656 -0.588530  0.863151
1 -0.203341  0.428186 -0.977210  0.226401 -0.124692  0.979151
2 -0.996631 -0.295198 -0.396213 -0.090265 -0.588913  0.690280
3  0.302886  0.060910 -0.078698 -0.481032  0.044984  1.244890
4 -0.921098  0.011266  0.203139 -0.179170 -0.107779  0.584488
```

Gambar 9. Hasil Proses *Embeddings* Menggunakan *IndoBERTweet*

DOI: https://doi.org/10.38204/tematik.v11i2.2096

3.4. *Dimensionality Reduction*

UMAP untuk *dimensionality reduction* disarankan dalam pendekatan *BERTopic*, meskipun teknik lain dapat diintegrasikan ke dalam tahap ini juga. *UMAP* merupakan algoritma yang didasarkan pada teknik dan ide pembelajaran *manifold* dari analisis data topologi [13]. Studi sebelumnya [14][15] mengkonfirmasi bahwa *UMAP* lebih optimal untuk kualitas pengelompokan daripada metode lain dari *dimensionality reduction*, seperti *t-distributed stochastic neighbor embedding (t-SNE)* dan *PCA*. *UMAP* efektif untuk menemukan penyematan dimensi rendah yang mempertahankan struktur topologi penting dari data.



Gambar 10. Hasil *Dimensionality Reduction* Menggunakan *UMAP*, *PCA*, dan t-*SNE* Secara Urut dari Kiri

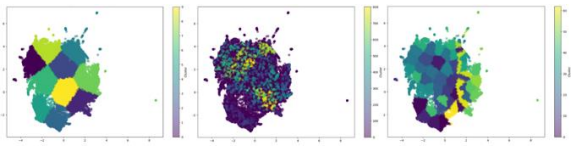
Gambar 10 menunjukkan *dimensionality reduction* menggunakan *UMAP* data tersebar dengan lebih banyak *cluster* yang berbeda, hal tersebut dikarenakan *UMAP* mempertahankan struktur lokal dan global dari data dengan baik, sehingga data yang serupa di ruang berdimensi tinggi tetap berdekatan di ruang dimensi rendah. Sedangkan *PCA* tidak mengelompokkan data menjadi *cluster* yang jelas seperti *UMAP*, dikarenakan *PCA* hanya mempertahankan struktur lokal data. *PCA* menunjukkan bahwa data tersebar dalam bentuk elips memanjang dengan lebih banyak data terpusat ditengah.

Sementara itu, hasil *dimensionality reduction* dengan *t-SNE* menghasilkan visualisasi dengan varian yang lebih besar pada satu sumbu, sehingga menyulitkan dalam identifikasi pola, hal tersebut diakibatkan *t-SNE* menangkap varians terbesar dalam data dan mempertahankan hanya struktur global data. Oleh karena itu, *t-SNE* sulit dalam mengidentifikasi *cluster-cluster* kecil dan *PCA* sulit dalam mengidentifikasi *cluster-cluster* besar, sehingga *dimensionality reduction* menggunakan *UMAP* pada penelitian ini lebih optimal dibandingkan dengan *PCA* dan *t-SNE*.

3.5. Clustering

Penggunaan *K-means* di *BERTopic* mencapai kinerja yang mirip dengan *HDBSCAN*, tetapi tanpa adanya *outliers*, sehingga memungkinkan pemodelan topik yang lebih akurat [8]. Algoritma *clustering K-means* menunjukkan tingkat konvergensi yang lebih cepat, efektif, dan efisien daripada *Birch* dalam pemodelan topik menggunakan *BERTopic* [16].

Gambar 11 menunjukkan *clustering* menggunakan *K-Means* mengelompokkan data berdasarkan banyak *cluster* yang telah ditetapkan sebelumnya parameter "n_clusters". *Cluster-cluster* yang dihasilkan *K-Means* direpresentasikan dengan warna-warna yang berbeda dengan besar *cluster* sebanding dengan jumlah kalimat dalam *cluster* tersebut. Letak *cluster* dalam *plot* mewakili rata-rata dari *embeddings* kalimat dalam *cluster* tersebut. Proses *clustering* dengan *K-Means* ini mengidentifikasi *cluster-cluster* kalimat yang mirip satu sama lain.



Gambar 11. Hasil *Clustering* Menggunakan *K-Means*, *HDBSCAN*, dan *Birch* Secara Urut dari Kiri

Sedangkan *HDBSCAN* mengidentifikasi *cluster* dengan ukuran yang bervariasi dikarenakan, jumlah *cluster* ditentukan otomatis berdasarkan kepadatan data, sehingga *cluster* yang dihasilkan tidak beraturan bentuk dibandingkan dengan *K-Means.* Pembentukan *cluster* dengan *HDBSCAN* juga membutuhkan kecepatan yang lambat yaitu sekitar 15 menit dalam proses *clustering.*

Sementara itu, hasil *clustering* dengan *Birch* mengidentifikasi jumlah *cluster* secara otomatis dengan mengeksplorasi *cluster* pada tingkat granularitas yang berbeda. Proses komputasi dalam *clustering* dilakukan, *Birch* memerlukan waktu yang cukup lama sekitar 10 menit. Pada kasus penelitian ini, *Birch* kurang optimal untuk digunakan dikarenakan, skalabilitas *Birch* lebih sesuai menganalisis koleksi teks yang besar.

3.6. Weighting Scheme

Penggunaan *c-TF-IDF* lebih baik daripada menggunakan algoritma lain untuk proses *weighting scheme* dalam *BERTopic* karena meningkatkan kemampuan dalam memproses data teks dengan distribusi yang tidak seimbang, serta meningkatkan akurasi klasifikasi teks [9]. Dalam representasi *TF-IDF* biasa, setiap kata memiliki bobot yang berat dalam mempertimbangkan istilah lokal: *Term Frequency (TF)* dan *Global term: Inverse Document Frequency (IDF).* Seperti yang ditunjukkan dalam persamaan (1), frekuensi istilah *tft,d* dihitung untuk istilah *t* dan dokumen *d* dan frekuensinya terbalik diperhitungkan sebagai logaritma dari jumlah dokumen N dalam dokumen atau *dataset* dibagi oleh jumlah total dokumen yang berisi istilah *t. (dft).*

$$W_{t,d} = tf_{t,d} \cdot \log\left(\frac{1+N}{df_t}\right) \quad (3)$$

Jadi versi modifikasi *TF-IDF* seperti yang ditunjukkan dalam Persamaan 3. Frekuensi istilah *tft,c* dihitung dengan mengikat semua dokumen ke dalam satu *cluster* dan menganggapnya sebagai satu dokumen *cluster c*. *IDF* juga dimodifikasi dan ditempatkan kembali oleh frekuensi *cluster* terbalik. Dihitung dengan mengambil logaritma dari jumlah rata-rata kata per *cluster* (A)

DOI: https://doi.org/10.38204/tematik.v11i2.2096

dibagi oleh frekuensi istilah *t* di seluruh *cluster*. *(tft)*, sehingga mengukur seberapa banyak informasi istilah memberikan ke *cluster* tertentu seperti pada Persamaan 4. Secara keseluruhan, pendekatan berbasis *TF-IDF* menghasilkan distribusi kata topik untuk setiap *cluster* dokumen

$$W_{t,c} = tf_{t,c} \cdot \log\left(\frac{1+A}{tf_t}\right) \quad (4)$$

3.7. Hasil Topik dengan *BERTopic*

Seusai proses *pre-processing* data telah selesai, maka *output* dari proses tersebut merupakan data akhir yang digunakan agar diproses ke dalam tahapan *BERTopic* untuk pembentukan topik. Tabel 2 menunjukkan hasil *representation tuning* dari *BERTopic* dengan memanfaatkan *OpenAI* untuk mempermudah pemahaman terhadap topik yang dihasilkan.

Tabel 2 menunjukkan kekhawatiran tentang keterjangkauan kendaraan listrik menjadi topik yang sering muncul, serta masyarakat masih merasa harga kendaraan listrik masih terlalu mahal. Selain itu, analisis biaya kendaraan listrik, yang mencakup biaya pembelian, perawatan, dan efisiensi bahan bakar, juga menjadi perhatian utama. Tantangan yang dihadapi oleh bangsa Indonesia dalam mengadopsi kendaraan listrik, termasuk infrastruktur, regulasi, dan kesadaran masyarakat, merupakan isu penting lainnya. Permintaan untuk kendaraan listrik yang sederhana dan terjangkau juga sering dibahas, menunjukkan kebutuhan masyarakat akan solusi mobilitas yang lebih ekonomis.

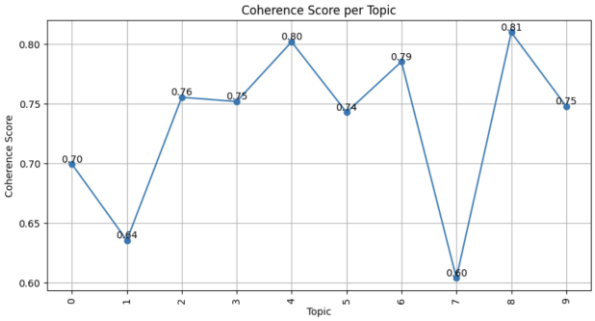
Tabel 2. Hasil *Representation Tuning* dari Pemodelan Topik Menggunakan Keseluruhan Data

No	*Count*	*Topic*
0	3091	*Electric Vehicle Affordability Concerns*
1	2526	*Electric Vehicle Cost Analysis*
2	2407	*Challenges Facing Indonesian Nation*
3	2386	*Affordable Simple Electric Vehicles*
4	2278	*Electric Vehicle Emission Testing*
5	2150	*Innovation for Affordable Mobility*
6	2119	*Challenges in Indonesian National Development*
7	2037	*Nationalistic Pride and Innovation*
8	1821	*Car Purchase Simplicity*
9	1765	*Electric Vehicle Technology Challenges*

3.8. Model Evaluation dengan *Coherence Score* dan *Topic Diversity*

Topik-topik yang terbentuk dievaluasi menggunakan *coherence score* dan *topic diversity*. Pada topik yang dihasilkan oleh data komentar tahun 2018 hingga 2024 memiliki nilai *diversity score* sebesar 0.95838, hal tersebut menunjukkan bahwa topik-topik yang dihasilkan oleh model memiliki keragaman yang tinggi dalam hal penggunaan kata-kata dan menangkap berbagai aspek dari dokumen teks dengan baik dan masing-masing topik memiliki fokus yang jelas.

Sedangkan *coherence score* yang dihasilkan sebesar 0.73321, dengan nilai koheren setiap topik tertera pada Gambar 12. Setiap topik menunjukkan nilai *coherence score* disekitar angka 0.6 hingga 1, dengan angka *coherence score* yang paling rendah pada topik ke-7 yaitu sebesar 0.60 dan angka *coherence score* paling tinggi pada topik ke-8 sebesar 0.81. Dengan demikian, topik ke-8 yaitu topik "Car Purchase Simplicity" yang terbentuk dari 1821 data komentar menjadi topik yang memiliki kesamaan semantik paling kuat.



Gambar 12. Representasi dari Variasi *Coherence Scores* dari Setiap Topik

Tabel 3. Perbandingan Algoritma *UMAP*, *K-Means*, *PCA*, *SVD*, *HDBSCAN*, dan *BIRCH* Menggunakan Dua Metrik Evaluasi

Algoritma *Dimensionality Reduction*	*Clustering*	*Coherence Score*	*Topic Diversity*
UMAP	*K-Means*	0.73883	0.95838
PCA	*K-Means*	0.70506	0.85917
t-SNE	*K-Means*	0.64477	0.89374
UMAP	*HDBSCAN*	0.53926	0.99996
UMAP	*BIRCH*	0.61141	0.97407

Tabel 3 menunjukkan pemilihan algoritma bergantung pada prioritas antara koherensi dan keberagaman topik yang dihasilkan. Algoritma yang dibandingkan merupakan algoritma yang digunakan pada proses *dimensionality reduction* dan *clustering*. *UMAP* dan *K-Means* memberikan hasil terbaik dalam hal keseimbangan antara koherensi dan keberagaman topik, *HDBSCAN* menghasilkan topik yang sangat beragam, tetapi dengan koherensi yang rendah, *PCA* dan *t-SNE* mencari keseimbangan antara koherensi dan keberagaman, dengan *PCA* sedikit lebih baik dalam hal koherensi, *BIRCH* menawarkan keberagaman topik yang sangat tinggi dengan koherensi yang lebih baik dibandingkan *HDBSCAN*, tetapi masih di bawah *UMAP & K-Means* dan *PCA*. Oleh karena itu, *UMAP* dan *K-Means* digunakan dalam proses *dimensionality reduction* dan *clustering* pada penelitian ini.

4. Kesimpulan

Dengan menggunakan *BERTopic*, topik yang dihasilkan mencapai nilai evaluasi yang baik berdasarkan kriteria kemiripan semantik yang tinggi serta, keragaman topik yang dihasilkan. Hasil evaluasi tersebut menunjukkan kinerja *BERTopic* yang unggul dalam menggunakan metrik pemodelan topik untuk menghasilkan topik-topik yang berbeda dan koheren.

DOI: https://doi.org/10.38204/tematik.v11i2.2096

Dari hasil analisis topik ditemukan kekhawatiran utama masyarakat terhadap keterjangkauan kendaraan listrik, serta pandangan masyarakat yang masih merasa harga kendaraan listrik terlalu mahal. Selain itu, analisis biaya kendaraan listrik, yang mencakup biaya pembelian, perawatan, dan efisiensi bahan bakar juga menjadi perhatian utama. Tantangan yang dihadapi oleh bangsa Indonesia dalam mengadopsi kendaraan listrik, seperti infrastruktur, regulasi, dan kesadaran masyarakat, menjadi isu penting lainnya. Permintaan untuk kendaraan listrik yang sederhana dan terjangkau juga sering dibahas, menunjukkan adanya kebutuhan masyarakat akan solusi mobilitas yang lebih ekonomis. Temuan ini dapat membantu pemerintah dalam mendengarkan opini publik dengan lebih baik, memahami dan mengatasi kekhawatiran, serta menghormati kehendak publik. Kemudian, dapat memberikan referensi penting untuk mengoptimalkan layanan publik dan merumuskan kebijakan yang masuk akal.

Daftar Rujukan

[1] ESDM, "Transisi Energi Bersih Melalui Kendaraan Bermotor Listrik," ESDM, 2020. https://www.esdm.go.id/id/berita-unit/direktorat-jenderal-ketenagalistrikan/transisi-energi-bersih-melalui-kendaraan-bermotor-listrik (accessed July. 17, 2024).

[2] V. Pirmana, A. S. Alisjahbana, A. A. Yusuf, R. Hoekstra, and A. Tukker, "Economic and environmental impact of electric vehicles production in Indonesia," *Clean Technologies and Environmental Policy*, vol. 25, Feb. 2023, doi: https://doi.org/10.1007/s10098-023-02475-6.

[3] M Askinatin, N Heldini, Y Supriyanto, None Saparudin, and N Ariyanto, "Analysis of market readiness for the safe use of electric vehicles in Indonesia post-pandemic era," *IOP Conference Series Earth and Environmental Science*, vol. 1267, no. 1, pp. 012042–012042, Dec. 2023, doi: https://doi.org/10.1088/1755-1315/1267/1/012042.

[4] Mardhi, Lu, "Pandangan Generasi Terkini Mengenai Kendaraan Listrik di Indonesia", Whiteboardjournal, 2023. (accessed by July. 17, 2024).

[5] Candra dan C, "Evaluasi hambatan untuk adopsi kendaraan listrik di Indonesia melalui pendekatan prioritas ordinal abu-abu", International Journal of Grey Systems, 2(1), 38-56, 2022

[6] B. Ogunleye, T. Maswera, L. Hirsch, J. Gaudoin, and T. Brunsdon, "Comparison of Topic Modelling Approaches in the Banking Context," *Applied Sciences*, vol. 13, no. 2, p. 797, Jan. 2023, doi: https://doi.org/10.3390/app13020797.

[7] Simanjuntak, K. A., Koyimatu, M., & Ervanisari, Y. P, "Analisis Perubahan Opini Publik Terhadap Kendaraan Listrik di Indonesia Melalui Komentar YouTube: Pendekatan Topic Modeling BERTopic", *Jurnal Inovasi Kewirausahaan*, *1*(3), 1-9, 2024, https://doi.org/10.37817/jurnalinovasikewirausahaan.v1i3

[8] Groot, M. Aliannejadi, and M. R. Haas, "Experiments on Generalizability of BERTopic on Multi-Domain Short Text," *arXiv (Cornell University)*, Jan. 2022, doi: https://doi.org/10.48550/arxiv.2212.08459.

[9] Z. Jiang, B. Gao, Y. He, Y. Han, P. Doyle, and Q. Zhu, "Text Classification Using Novel Term Weighting Scheme-Based Improved TF-IDF for Internet Media Reports," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–30, Mar. 2021, doi: https://doi.org/10.1155/2021/6619088.

[10] H. P. Suresha and K. Kumar Tiwari, "Topic Modeling and Sentiment Analysis of Electric Vehicles of Twitter Data," *Asian Journal of Research in Computer Science*, pp. 13–29, Oct. 2021, doi: https://doi.org/10.9734/ajrcos/2021/v12i230278.

[11] A. Uteuov, "Topic model for online communities' interests prediction," *Procedia Computer Science*, vol. 156, pp. 204–213, 2019, doi: https://doi.org/10.1016/j.procs.2019.08.196.

[12] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, doi: https://doi.org/10.18653/v1/2021.emnlp-main.833.

[13] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform Manifold Approximation and Projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, Sep. 2018, doi: https://doi.org/10.21105/joss.00861.

[14] K. Kukushkin, Y. Ryabov, and A. Borovkov, "Digital Twins: A Systematic Literature Review Based on Data Analysis and Topic Modeling," *Data*, vol. 7, no. 12, p. 173, Nov. 2022, doi: https://doi.org/10.3390/data7120173.

[15] Y. Yang *et al.*, "Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data," *Cell Reports*, vol. 36, no. 4, p. 109442, Jul. 2021, doi: https://doi.org/10.1016/j.celrep.2021.109442.

[16] F. Nie, Z. Li, R. Wang, and X. Li, "An Effective and Efficient Algorithm for K-Means Clustering With New Formulation," vol. 35, no. 4, pp. 3433–3443, Jan. 2022, doi: https://doi.org/10.1109/tkde.2022.3155450.

DOI: https://doi.org/10.38204/tematik.v11i2.2096