10.37817/jurnalinovasikewirausahaan.v1i3
E-ISSN : 3046-6288
P-ISSN: 3046-6296

Analisis Perubahan Opini Publik Terhadap Kendaraan Listrik di Indonesia Melalui Komentar *YouTube:* Pendekatan *Topic Modeling BERTopic*

[1]Kristine Angelina Simanjuntak, [2]Muhamad Koyimatu, [3]Yolla Putri Ervanisari
[1]Ilmu Komputer, Universitas Pertamina, Jakarta Selatan
[2]Ilmu Komputer, Universitas Pertamina, Jakarta Selatan
[3]Ilmu Komputer, Universitas Pertamina, Jakarta Selatan

E-mail: [1]105220009@student.universitaspertamina.ac.id, [2]koyimatu@universitaspertamina.ac.id, [3]105220049@student.universitaspertamina.ac.id

ABSTRAK

Penelitian ini membahas bagaimana sebuah perubahan opini publik terhadap kendaraan listrik di Indonesia dapat dianalisis melalui komentar *YouTube* dengan menggunakan pendekatan *topic modeling* menggunakan algoritma *BERTopic*. Terdapatnya kontroversial terhadap masuknya kendaraan listrik di Indonesia yang memicu perdebatan tentang dampak ekonomi dan lingkungan. Menganalisis perubahan opini publik terhadap kendaraan listrik di Indonesia melalui komentar pada *YouTube* menggunakan pendekatan *topic modeling BERTopic,* serta mengevaluasi hasil analisis tersebut menjadi tujuan dari penelitian. Judul artikel ini adalah "Analisis Perubahan Opini Publik Terhadap Kendaraan Listrik di Indonesia Melalui Komentar YouTube: Pendekatan Topic Modeling BERTopic". Metode yang digunakan algoritma *BERTopic* mulai dari melakukan *embedding* hingga *representation tuning* dari topik dihasilkan. Dalam mengumpulkan data, digunakan teknik *web crawling* untuk mengambil data komentar publik dari berbagai video dengan fokus kepada komentar publik yang memberikan opininya terhadap kendaraan listrik di Indonesia. Hasil penelitian ini memperoleh pengumpulan data komentar sebanyak 138.662 data dengan data akhir yang akan digunakan setelah melalui tahapan *preprocessing* sebanyak 49.104 data. Kemudian data akhir tersebut menghasilkan sepuluh topik yang merepresentasikan data komentar dengan topik "Electric vs Gasoline Car Economics" menjadi topik yang merepresentasikan 6.856 data komentar dari total 49.104 total data. Komentar dari pengguna *YouTube* yang diberikan masih terfokus membahas seputar perbandingan antara dampak ekonomi menggunakan kendaraan listrik dengan kendaraan berbahan bakar fosil.

Kata kunci : ***BERTopic, Crawling, Kendaraan Listrik, Opini Publik, Topic Modeling, YouTube***

ABSTRACT

This study discusses how a change in public opinion on electric vehicles in Indonesia can be analyzed through YouTube comments using a topic modeling approach using the BERTopic algorithm. There was controversy over the introduction of electric vehicles in Indonesia which sparked a debate about the economic and environmental impact. Analysing the change in public opinion on electric vehicles in Indonesia through comments on YouTube using the BERTopic modeling topic approach, as well as evaluating the results of such analysis as the objective of the research. The title of this article is "Analysis of Public Opinion Changes to Electric Vehicles in Indonesia Through YouTube Comments: BERTopic Topic Modeling Approaches". The method used by the BERTopic algorithm ranges from embedding to tuning representation of the topic produced. In collecting data, web crawling techniques are used to retrieve public comment data from various videos with a focus on public comments that give their opinions about electric vehicles in Indonesia. The results of this study resulted in the collection of 138,662 comments data with final data to be used after a preprocessing phase of 49.104 data. Then the final data produced ten topics representing comment data with the topic "Electric vs Gasoline Car Economics" to the topic representing 6,856 comments data out of a total of 49,104 total data. The comments given by YouTube users are still focused around the comparison of the economic impact of using electric vehicles with fossil-fueled vehicles.

Keywords : BERTopic, Crawling, Electric Vehicles, Public Opinion, Topic Modeling, YouTube

E-ISSN : 3046-6288
P-ISSN: 3046-6296
10.37817/jurnalinovasikewirausahaan.v1i3

1. PENDAHULUAN

Kedatangan kendaraan listrik atau yang lebih sering dikenal sebagai *Electric Vehicle (EV)* di Indonesia telah memicu perdebatan tentang dampak ekonomi dan lingkungan. Sementara *EV* yang didukung oleh listrik berbasis batubara dapat mengurangi emisi seperti PM2.5, dapat meningkatkan emisi CO2, menekankan kebutuhan untuk analisis komprehensif dari kompromi yang terlibat dalam transisi ke *EV*. (Pirmana et al., 2023). Selain itu, status kesiapan *EV* di Indonesia, seperti yang dirasakan oleh berbagai *stakeholder*, mempengaruhi kecepatan adopsi. Namun, tantangan seperti fasilitas pengisian yang terbatas dan harga tinggi mencegah konsumen potensial dari merangkul EV (Askinatin, 2023).

Mengidentifikasi dan mengatasi hambatan untuk adopsi *EV*, seperti infrastruktur pengisian daya yang terbatas dan biaya tinggi, sangat penting untuk mempercepat transisi. (Candra, 2022; Askinatin, 2023). Memahami persepsi masyarakat dapat membantu dalam menentukan strategi pemasaran dan mengoptimalkan potensi kendaraan listrik di pasar Indonesia. Dengan demikian, aspek sosial dan penerimaan masyarakat menjadi kunci dalam mempercepat adopsi kendaraan listrik. Pemerintah Indonesia juga terlibat aktif dalam mendorong konversi kendaraan berbasis listrik, sebagaimana yang dibahas oleh (Aziz et al., 2020).

Analisis opini publik sangat penting ketika mempertimbangkan pengenalan *EV* di Indonesia karena sifat kontroversial dari transisi ini. Sentimen publik memainkan peran penting dalam membentuk kebijakan dan inisiatif yang terkait dengan adopsi *EV*. Memahami pandangan, kekhawatiran, dan preferensi masyarakat tentang *EV* dapat membantu pembuat kebijakan menyesuaikan strategi yang mengatasi hambatan potensial dan mempromosikan penerimaan. (Maghfiroh et al., 2021). Sebagai kesimpulan, analisis opini publik sangat penting untuk integrasi *EV* yang sukses di Indonesia. Opini tersebut disampaikan oleh masyarakat melalui berbagai media sosial, salah satunya *YouTube*.

YouTube merupakan media sosial yang berfokus menyediakan berbagai konten video yang telah digunakan sebanyak 170 juta dari 181,9 juta total pengguna internet berusia 16-64 tahun, dengan 179,1 juta orang di Indonesia telah menggunakan *YouTube* untuk menonton video *online* (Krisna, 2021). Melalui video *YouTube*, masyarakat menyampaikan opininya terkait dengan video yang mereka nonton menggunakan fitur komentar. Opini-opini yang dituangkan oleh masyarakat berperan penting dalam pengembangan kebijakan pemerintah Indonesia. Untuk mengidentifikasi opini masyarakat, penggunaan *topic modeling* sangat bermanfaat dalam mengelompokkan opini-opini yang berbeda ke dalam topik-topik tertentu (Wu et al., 2021).

Topic modeling merupakan sebuah metode untuk mengidentifikasi topik-topik dari sekelompok dokumen yang tidak terstruktur (Sirkis & Maitland, 2022). Salah satu teknik *topic modeling* yang digunakan adalah *BERTopic*. *BERTopic* merupakan teknik pemodelan yang menggunakan *embeddings BERT* dan *c-TF-IDF* untuk membuat topik (Grootendorst, 2022). Pada penelitian Sharifian et al. (2022) menunjukkan bahwa performa *BERTopic* lebih bagus dibanding dengan teknik pemodel topik lainnya, dalam hal koherensi dan keragaman topik yang dihasilkan 18 kali lebih cepat.

Penelitian Suresha (2021) yang berjudul "Topic Modeling and Sentiment Analysis of Electric Vehicles of Twitter Data" 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. Dalam penelitian Ogunleye (2023) 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 (Ogunleye et al., 2023).

Oleh karena itu, dalam penelitian ini akan 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. LANDASAN TEORI

2.1 Topic Modeling

Topic Modeling adalah teknik pembelajaran mesin yang banyak digunakan di berbagai bidang seperti analisis literatur, penelitian komunikasi, dan penelitian medis. Teknik ini melibatkan pengungkapan struktur tematik laten dari koleksi dokumen yang besar (Cao et al., 2022). *Topic modeling* memungkinkan untuk menganalisis dan memahami opini yang berkembang di dalam komunitas *online* tertentu dari waktu ke waktu. Selain itu, *topic modeling* telah diterapkan pada unggahan media sosial untuk mengidentifikasi topik dan tema yang sedang tren, memberikan cara untuk mengeksplorasi diskusi dan minat dalam percakapan *online* (Liao & Liu, 2023). Dengan mengekstraksi topik dari komentar media sosial, peneliti dapat memperoleh wawasan tentang konten yang dibagikan dan didiskusikan oleh pengguna.

2.2 BERTopic

BERTopic dapat dilihat sebagai urutan langkah-langkah untuk membuat representasi topik. Proses ini melibatkan pengubahan dokumen menjadi *word embedding representations* menggunakan model bahasa yang telah dilatih sebelumnya *(pre-trained language model)*, melakukan *dimensionality reduction* melalui teknik seperti *UMAP*, mengelompokkan teks ke dalam kelompok-kelompok yang memiliki arti yang sama menggunakan algoritma seperti *K-Means*, dan akhirnya, memanfaatkan versi *TF-IDF* berbasis kelas untuk mengekstrak representasi topik dari setiap topik (Scarpino dkk., 2022). Selain itu, *BERTopic* juga melibatkan pemanfaatan *word embeddings*, *feature reduction*, dan algoritma pengelompokan dalam kerangka kerjanya. Algoritma ini membuat *document embeddings* menggunakan model *transformer-based language model*, mengurangi *dimensionality* dari *embeddings*, dan membentuk cluster yang secara semantik mirip untuk mengekstrak representasi topik (Huang, 2023).

2.2.1 IndoBERTweet

IndoBERTweet adalah model *pre-trained* khusus yang dirancang untuk analisis teks dalam bahasa Indonesia. Model ini didasarkan pada model *BERT (Bidirectional Encoder Representations from Transformers)*, sebuah model *Natural Language Processing (NLP)*. *IndoBERTweet* membuat *topic modeling* dalam bahasa Indonesia telah mengalami kemajuan yang signifikan dengan diperkenalkannya *IndoBERTweet*, sebuah model bahasa yang sudah terlatih yang secara khusus dirancang untuk analisis teks bahasa Indonesia (Koto dkk. 2020). Kesimpulannya, *IndoBERTweet* merupakan model yang lebih terfokus pada yang cenderung lebih singkat, informal, dan mungkin mengandung slang, singkatan, atau istilah yang unik.

2.2.2 UMAP

UMAP merupakan singkatan dari *Uniform Manifold Approximation and Projection* adalah teknik reduksi dimensi yang biasa digunakan dalam pemodelan topik. Teknik ini bekerja dengan memetakan data berdimensi tinggi ke dalam ruang berdimensi lebih rendah dengan tetap mempertahankan struktur dan hubungan yang mendasari data. (Ordun et al., 2020). Dalam konteks pemodelan topik, *UMAP* digunakan untuk mengurangi dimensionalitas *embedding* dokumen. Kemampuan *UMAP* dalam menangani data berdimensi tinggi semakin ditekankan oleh kemampuannya untuk membedakan dengan jelas *cluster-cluster* yang berdekatan sambil mempertahankan struktur data secara keseluruhan (Sakaue et al., 2020).

2.2.3 K-Means

Algoritma *cluster* yang paling sering digunakan *disebut k-means cluster*. Beberapa algoritma *cluster* membutuhkan jumlah *cluster* yang harus dipilih terlebih dahulu. Ini adalah kasus dengan *k-means clustering*. Seringkali mencari tahu jumlah *cluster* bisa menjadi tugas yang paling sulit dalam *cluster* (Eskonen, J, 2022). Menemukan cluster menggunakan *K-Means* yang memungkinkan untuk memilih berapa banyak *cluster* yang diinginkan dan memaksa setiap titik untuk berada di dalam

E-ISSN : 3046-6288
P-ISSN: 3046-6296
10.37817/jurnalinovasikewirausahaan.v1i3

cluster. Oleh karena itu, tidak akan ada *outlier* yang diciptakan.

2.2.4 CountVectorizer

CountVectorizer adalah alat penting dalam pemodelan topik yang digunakan untuk melakukan prapemrosesan data teks sebelum menerapkan algoritme pembelajaran mesin untuk mengekstrak topik dari kumpulan dokumen. Alat ini biasanya digunakan untuk mengubah data teks menjadi format numerik yang dapat digunakan oleh model pembelajaran mesin (Rustam et al., 2021). Dengan menggunakan *CountVectorizer, stop words* dapat dihilangkan untuk meningkatkan kualitas model topik yang dihasilkan (Verbeij et al., 2022). Pemodelan topik, sebuah teknik pembelajaran mesin, melibatkan ekstraksi topik laten dari sekumpulan dokumen tanpa memerlukan kamus atau aturan interpretasi yang telah ditetapkan sebelumnya (Storopoli, 2019). Hal ini memungkinkan identifikasi kelompok kata yang muncul bersama yang mewakili konsep tingkat tinggi dalam data teks (Storopoli, 2019).

2.2.5 c-TF-IDF

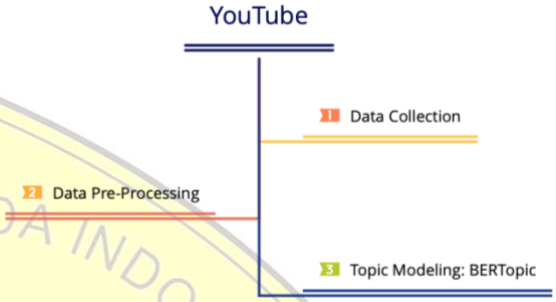
c-TF-IDF adalah versi yang disempurnakan dari algoritme *TF-IDF (Term Frequency-Inverse Document Frequency)* tradisional yang digunakan dalam pemodelan topik. Algoritme *c-TF-IDF* menggabungkan statistik *chi-square* untuk meningkatkan identifikasi istilah-istilah penting dalam kumpulan dokumen. Dengan menggabungkan *TF-IDF* dengan statistik *chi-square*, *c-TF-IDF* memberikan bobot pada istilah berdasarkan frekuensi mereka dalam dokumen dan signifikansinya dalam membedakan dokumen tersebut dengan dokumen lain dalam korpus (Liu, 2023). Dalam konteks pemodelan topik, *c-TF-IDF* memainkan peran penting dalam mengidentifikasi topik utama dalam kumpulan dokumen. Dengan memanfaatkan statistik *chi-square*, *c-TF-IDF* dapat menangkap karakteristik unik dari setiap dokumen dengan lebih baik terutama istilah-istilah yang tidak hanya sering muncul tetapi juga khas.

2.2.6 OpenAI

OpenAI adalah kecerdasan buatan yang mengembangkan model bahasa *GPT-3* dan *ChatGPT*. *ChatGPT*, sebuah model bahasa yang dibuat oleh *OpenAI*, merupakan transformator generatif yang telah dilatih sebelumnya yang dapat menghasilkan respons teks yang mirip manusia berdasarkan petunjuk atau konteks yang diberikan (Ho, 2023; Curtis, 2023). Berbeda dengan model bahasa searah *(unidirectional language model)*, *ChatGPT* mempertimbangkan konteks sebelum dan sesudahnya, sehingga sangat cocok untuk tugas-tugas seperti pemodelan topik (Meier, 2023).

3. METODOLOGI

Metodologi penelitian yang dilakukan pada penelitian akan dijelaskan pada Gambar 1.



Gambar 1. *Flowchart* Metodologi Penelitian

3.1 Data Collection

Penelitian ini melakukan pengumpulan data dengan metode *web crawling.* Data yang akan dikumpulkan berupa komen publik dari berbagai video *youtube* yang muncul sesuai dengan beberapa kata kunci yang telah ditentukan. Data yang diambil berupa *"*author" menunjukkan pemilik akun, "updated_at" menunjukkan waktu komen dikirimkan, "like_count" menunjukkan jumlah *like* dari komen, "text" menunjukkann isi komen, "video_id" menunjukkan *id* video dari komen tersebut, dan "public" menunjukkan komen tersebut dapat diakses secara *public*. *Crawling data* dilakukan dengan mengambil *id* dari beberapa video yang muncul dari hasil pencarian dengan menggunakan kata kunci dengan mem*filter* berdasarkan *view count.* Pengumpulan data video ini megambil dari tahun 2018 hingga tahun 2024, dengan detail banyaknya video pada tertera pada Gambar dibawah.

3.2 Data Pre-Processing

Setelah proses *crawling* telah dilakukan, *raw dataset* atau kumpulan data mentah yang tidak memiliki format yang teratur akan 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 dalam penelitian ini antara lain:

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

2. *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.

3.3 Topic Modeling: BERTopic

3.3.1 Embedding

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*.

3.3.2 Dimensionality Reduction

Dalam *BERTopic*, algoritma *dimensionality reduction* digunakan untuk mengurangi jumlah dimensi atau fitur dalam sekumpulan data. Ini dilakukan 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.

3.3.3 Clustering

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.

3.3.4 Vectorizer

Tahapan *vectorizer* dengan menggunakan *CountVectorizer* dapat dilakukan beberapa hal seperti:

1. Menghapus *stopwords* kata-kata umum yang tidak memberi makna seperti “dan”, “atau”, dan sebagainya.

2. Mengabaikan kata-kata yang jarang muncul dan tidak relevan.

3. Tokenisasi, memecah teks menjadi kata-kata atau token dengan memisahkan teks menjadi unit-unit yang lebih kecil seperti berdasarkan spasi atau pola tertentu.

3.3.5 Weighting Scheme

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.

3.3.6 Representation Tuning

Proses mengubah representasi topik yang dihasilkan oleh model *BERTopic*. *Representation tuning* ini dapat dilakukan dengan menggunakan beberapa model yang telah terimplementasikan dalam *BERTopic*, seperti *OpenAI, KeyBERTInspired*, dan lain-lain. *OpenAI* adalah model yang menggunakan *API OpenAI* untuk mengelompokkan topik. *KeyBERTInspired* adalah model yang menggunakan algoritma *KeyBERT* untuk membantu mengelompokkan topik. Ini dapat digunakan untuk memperbaiki kualitas topik dan memperingkatkan koherensi topik.

E-ISSN : 3046-6288
P-ISSN: 3046-6296
10.37817/jurnalinovasikewirausahaan.v1i3

4. HASIL DAN PEMBAHASAN

4.1 Data Collection

Proses pengumpulan data dengan metode *crawling* seperti yang telah dijelaskan pada tahapan metode penelitian, sumber data penelitian yang diambil berasal dari media sosial *YouTube* yang mengumpulan data opini atau komentar masyarakat selama tujuh tahun mulai dari tahun 2018 hingga 2024. Hasil *crawling* memperoleh 138.662 data komentar.

	author	text
0	@baiqwarni1831	Mau. Cuma masih liat2 isi rekening
1	@fandyahmadissuma1669	Gimana mesannya
2	@user-fd2ft1ob9v	Lucu donk pengen punya harga brp tuh
3	@seoranghamba7	berapa ya hrg nya
4	@raaldzikrogroup7822	Umurannya berapa itu kok kecil seksli..brapa h...
...	...	...
305	@insaniyahaja1049	Kpn mau dijual di indonesia ya?
306	@slametmd2	Hahaha...mana ada yang murah kalo udah dijual ...
307	@winasisprayitno6968	Masalahnya di charge baterai...itu berarti rum...
308	@herrynike953	Luar biasa menarik ya? Tp daya listrik di ruma...
309	@bangngadatu	Harga baru menyesuaikan kendaraan lain ikut na...

138662 rows × 6 columns

Gambar 2. Hasil *Data Crawling* dari *YouTube Video*

4.2 Data Pre-Processing

Setelah proses *crawling* data berhasil, selanjutnya hasil tersebut disimpan kedalam sebuah *data frame* untuk memudahkan proses pembersihan data. Beberapa tahapan *pre-processing* yang dilakukan dalam penelitian ini antara lain:

1. *Indonesian Language Data Filter*

Sebagai tahap awal, *data frame* yang masih tidak terstruktur dari hasil *crawling* akan di *filter* hanya komentar yang menggunakan bahasa berbahasa Indonesia dengan hasil akhir jumlah data yang diperoleh adalah 82.074 data, dengan kata lain ada sebanyak 56.588 data yang tidak berbahasa Indonesia.

	author	text
1	@fandyahmadissuma1669	Gimana mesannya
2	@user-fd2ft1ob9v	Lucu donk pengen punya harga brp tuh
3	@seoranghamba7	berapa ya hrg nya
4	@raaldzikrogroup7822	Umurannya berapa itu kok kecil seksli..brapa h...
5	@user-ij9mt9eg1o	Pengen beli aku tp dmn ya
...	...	...
138650	@bennyhendra3722	sdh naik harga ya dari 75 juta jadi 90 juta
138653	@fauziyahhusni7724	Ya elah udah semangat y nonton harga murah,eh ...
138657	@insaniyahaja1049	Kpn mau dijual di indonesia ya?
138658	@slametmd2	Hahaha...mana ada yang murah kalo udah dijual ...
138659	@winasisprayitno6968	Masalahnya di charge baterai...itu berarti rum...

82074 rows × 6 columns

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

2. *Word Distribution Filter*

Pada bagian ini pertama akan dilakukan proses menyeleksi komentar yang hanya memiliki minimal delapan kata per komen, dengan hasil akhir dari jumlah data yang diperoleh adalah sebanyak 49.104 data yang berarti ada sebanyak 32.970 data yang tidak memenuhi syarat penyeleksian jumlah kata.

	author	text
5	@user-wp8dc3ss2f	Alah lek cumak segituu mending motorrr bisa ng...
7	@suyuthiidrismarrappa	sayang sekali Mas Ridwan Hanif tidak membocork...
10	@gamerindo2295	Di pake ibu ibu dah kacau dah perjalanan indon...
14	@yuyunyuningsih462	keren di tambah bahasa penyampaian nya enak ba...
16	@dojowarrior7811	Pernah mogok pake ini . Mesin mati total. Sete...
...	...	...
82067	@zen.quotes	hmmmmm bisa speed up to 155km/h dan range 300k...
82069	@bennyhendra3722	sdh naik harga ya dari 75 juta jadi 90 juta
82070	@fauziyahhusni7724	Ya elah udah semangat y nonton harga murah,eh ...
82072	@slametmd2	Hahaha...mana ada yang murah kalo udah dijual ...
82073	@winasisprayitno6968	Masalahnya di charge baterai...itu berarti rum...

49104 rows × 6 columns

Gambar 4. Hasil *Pre-Processing* dengan Menggunakan *Word Distribution Filter*

Pemilihan jumlah kata yang ingin diseleksi pada setiap komen didasarkan pada perhitungan *word distribution* untuk melihat distribusi kata dalam setiap kalimat atau data komentar. Hal tersebut diperlihatkan pada Gambar 5, yang menunjukkan bahwa pada kuartil pertama dari data (25% bagian) memiliki data komentar yang mengandung delapan kata, sehingga data komentar yang akan diambil hanya data komentar yang mengandung minimal delapan kata dalam suatu kalimat.

E-ISSN : 3046-6288
P-ISSN: 3046-6296

	like_count	word_distribution
count	65891.000000	65891.000000
mean	1.767525	21.774931
std	19.428433	30.195615
min	0.000000	1.000000
25%	0.000000	8.000000
50%	0.000000	14.000000
75%	0.000000	25.000000
max	1991.000000	1738.000000

Gambar 5. Distribusi Kata

4.3 Representation Tuning

Setelah proses *pre-processing* data telah selesai, maka *output* dari proses tersebut merupakan data akhir yang akan digunakan agar diproses ke dalam tahapan *BERTopic* untuk pembentukan topik. Pada Tabel 1 menunjukkan hasil *representation tuning* dari *BERTopic* dengan memanfaatkan *OpenAI* dengan menampilkan lima topik yang tertera pada Tabel 1.

Tabel 1. Representasi Topik yang Dihasilkan

Topic	*Count*	*OpenAI (Topic)*
1	6856	*Electric vs Gasoline Car Economics*
2	5800	Mobil Subsidi Beli YG
3	5380	*Future of Affordable Electric Vehicles*
4	5174	*Pricing of Vehicle in Indonesia*
5	4735	*Indonesian Electric Vehicle Subsidies*

5. KESIMPULAN

Perubahan opini publik terhadap kendaraan listrik di Indonesia melalui pemodelan *BERTopic* menggunakan metode *topic modeling* antara lain diperoleh lima topik dengan topik "Electric vs Gasoline Car Economics" menjadi inti utama dari opini publik yang paling banyak dibahas oleh publik selama tujuh tahun mulai dari tahun 2018 hingga 2024 dengan memperoleh nilai "Count" sebanyak 6856 yang berarti ada sebanyak 6856 data komentar yang memiliki inti pembahasan yang mendeskripsikan topik tersebut dalam video yang membahas tentang kendaraan listrik di Indonesia. Melalui topik tersebut dapat disimpulkan bahwa komentar *YouTube* masih terfokus dalam membandingkan harga dalam pemakaian kendaraan listrik dengan kendaraan berbahan bakar fosil. Selain itu, memanfaatkan *OpenAI* dalam merepresentasikan topik sangat bermanfaat dalam memahami hasil pemodelan topik, dikarenakan *OpenAI* meniru bahasa manusia, sehingga topik yang dihasilkan dapat dimengerti dengan mudah oleh orang awam.

DAFTAR PUSTAKA

Askinatin, M., Heldini, N., Supriyanto, Y., & Ariyanto, N. (2023, December). Analysis of market readiness for the safe use of electric vehicles in Indonesia post-pandemic era. In *IOP Conference Series: Earth and Environmental Science* (Vol. 1267, No. 1, p. 012042). IOP Publishing.

Aziz, M., Marcellino, Y., Rizki, I., Ikhwanuddin, S., & Simatupang, J. (2020). Studi analisis perkembangan teknologi dan dukungan pemerintah indonesia terkait mobil listrik. Tesla Jurnal Teknik Elektro, 22(1), 45. https://doi.org/10.24912/tesla.v22i1.7898

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

Cao, Q., Cheng, X., & Liao, S. S. (2022). A comparison study of topic modeling-based literature analysis by using full texts and abstracts of scientific articles: a case of covid-19 research. Library Hi Tech, 41(2), 543-569.

Curtis, N. (2023). To chatgpt or not to chatgpt? the impact of artificial intelligence on academic publishing. The Pediatric

Infectious Disease Journal, 42(4), 275-275.

Eskonen, J. (2022). *Dynamic Topic Modeling and Clustering: Dynamic Topic Modeling and Clustering of Occupational Health and Safety Publications* (Master's thesis).

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv*, abs/2203.05794.

Ho, S. (2023). Circle packing charts generated by chatgpt to identify the characteristics of articles by anesthesiology authors in 2022: bibliometric analysis. Medicine, 102(50), e34511.

Huang, S. and Cole, J. M. (2023). Chemdatawriter: a transformer-based toolkit for auto-generating books that summarise research. Digital Discovery, 2(6), 1710-1720.

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). Indolem and indobert: a benchmark dataset and pre-trained language model for indonesian nlp. Proceedings of the 28th International Conference on Computational Linguistics.

Krisna, I. (2021). Analisis Comments To Subscriber Ratio Youtube Pada 10 Youtuber Indonesia Dengan Penghasilan Paling Fantastis.

Liao, Z. and Liu, X. (2023). Trending topics and themes in environmental innovation research based on topic modeling. Sustainable Development, 31(4), 2967-2978.

Liu, R. (2023). A detection methodology for sql injection attacks based on the tf-idf-chi algorithm.

Maghfiroh, M., Pandyaswargo, A., dan Onoda, H. (2021). Status ketersediaan saat ini kendaraan listrik di Indonesia: persepsi multistakeholder. Keberlanjutan, 13(23), 13177

Meier, F. (2023). Navigating the frontier of synthetic biology: an ai-driven analytics platform for exploring research trends and relationships. Acs Synthetic Biology, 12(11), 3229-3241.

Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., & Brunsdon, T. (2023). Comparison of Topic Modelling Approaches in the Banking Context. *Applied Sciences*.

Ordun, C., Purushotham, S., & Raff, E. (2020). Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs.

Pirmana, V., Alisjahbana, A, Yusuf, A., Hoekstra, R., & Tukker, A. (2023). dampak ekonomi dan lingkungan dari produksi kendaraan listrik di Indonesia. Teknologi Bersih dan Kebijakan Lingkungan, 25(6), 1871-1885.

Rustam, F., Ashraf, I., Shafique, R., Mehmood, A., Ullah, S., & Choi, G. (2021). Review prognosis system to predict employees job satisfaction using deep neural network. Computational Intelligence, 37(2), 924-950.

Sakaue, S., Hirata, J., Kanai, M., Suzuki, K., Akiyama, M., Too, C. L., … & Okada, Y. (2020). Dimensionality reduction reveals fine-scale structure in the japanese population with consequences for polygenic risk prediction. Nature Communications, 11(1).

Scarpino, I., Zucco, C., Vallelunga, R., Luzza, F., & Cannataro, M. (2022). Investigating topic modeling techniques to extract meaningful insights in italian long covid narration. BioTech, 11(3), 41.

Sharifian-Attar, V., De, S., Jabbari, S., Li, J., Moss, H., & Johnson, J. (2022). Analysing Longitudinal Social Science Questionnaires: Topic modelling with BERT-based Embeddings. *2022 IEEE International Conference on Big Data (Big Data)*, 5558-5567.

Sirkis, T. and Maitland, S. (2022). Monitoring real-time junior doctor sentiment from comments on a public social media platform: a retrospective observational study. Postgraduate Medical Journal, 99(1171), 423-427.

Storopoli, J. (2019). Topic modeling: how and why to use in management research. Revista Ibero-Americana De Estratégia, 18(3), 316-338.

Suresha, H. P., & Tiwari, K. K. (2021). Topic Modeling and Sentiment Analysis of Electric Vehicles of Twitter Data. *Asian J. Res. Comput. Sci*, 13-29.

Tang, Z., Pan, X., & Gu, Z. (2024). Analyzing public demands on China's online government inquiry platform: A BERTopic-Based topic modeling study. *Plosone*, *19*(2), e0296855

Verbeij, T., Beyens, I., Trilling, D., & Valkenburg, P. (2022). Happiness and sadness in adolescents' instagram direct messaging: a neural topic modelling approach.

Wu, B., Yuan, T., Qi, Y., & Dong, M. (2021). Public opinion dissemination with incomplete information on social network: a study based on the infectious diseases model and game theory. *Complex System Modeling and Simulation*, *1*(2), 109-121.