Analisis Emosi Publik Terhadap Isu Operasi Tangkap Tangan Gubernur Riau Abdul Wahid Pada Platform Youtube Sebagai Dasar Evaluasi Respon Dan Kebijakan Publik

Laporan Akhir ini Disusun untuk Memenuhi Ujian Akhir Semester

Mata Kuliah Pemrosesan Teks

2024B - Kelompok 2

Nama Anggota:

1. Muhammad Fahri Novrian (1314623014)
2. Christine Aprilia Putri (240311554046)
3. Ayu Wulan Anggraeni Putri (24031554177)
4. Fridania Nisa Calita (24031554208)

Dosen Mata Kuliah:

Ulfa Siti Nuraini, S.Stat., M.Stat.

Dr. Dian Handayani, M.Si.

Sains Data, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Surabaya

S1 Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta

2025

DAFTAR ISI

DAFTAR ISI ii
BAB 1 PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Tujuan 1
1.3 Manfaat 2
BAB 2 TINJAUAN PUSTAKA 3
2.1 Landasan Teori 3
BAB 3 METODOLOGI 5
3.1 Diagram Alur dan Tahapan Analisis 5
BAB 4 HASIL DAN PEMBAHASAN 9
4.1 Data Penelitian 9
4.2 Evaluasi Model Labeling Emosi 10
4.3 Analisis Emosi 12
4.4 Feature Engineering 14
4.5 Pemodelan LDA 14
4.6 Analisis Topik dan Emosi 18
BAB 5 KESIMPULAN 22
DAFTAR PUSTAKA 23

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi dan media sosial telah mengubah cara masyarakat menyampaikan opini, kritik, dan emosi terhadap berbagai isu sosial maupun politik. Salah satu platform yang digunakan masyarakat Indonesia untuk mengekspresikan pendapat adalah *YouTube*. Melalui kolom komentar, publik secara bebas memberikan tanggapan terhadap berbagai peristiwa yang terjadi, termasuk kasus-kasus korupsi dan penegakan hukum.

Salah satu isu yang menarik perhatian publik saat ini adalah Operasi Tangkap Tangan (OTT) terhadap Abdul Wahid yang merupakan Gubernur Riau oleh Komisi Pemberantasan Korupsi (KPK). Peristiwa ini menimbulkan beragam reaksi dari masyarakat. Analisis terhadap komentar publik di media sosial dapat memberikan informasi terkait persepsi dan emosi masyarakat terhadap isu tersebut.

Dalam proyek ini, dilakukan pendekatan hybrid yang menggabungkan analisis emosi dan analisis topik. Untuk proses labeling emosi, digunakan dua metode berbeda yaitu IndoBERT dan EmoLex. Kedua metode ini digunakan untuk melakukan pelabelan emosi pada data komentar, kemudian hasilnya dibandingkan untuk menentukan metode mana yang memberikan akurasi lebih baik. Selain itu, metode *Latent Dirichlet Allocation* (LDA) digunakan untuk mengetahui kepada siapa atau pada topik apa emosi tersebut ditujukan, misalnya apakah emosi marah lebih sering diarahkan kepada Gubernur, KPK, atau pemerintah. Kombinasi antara analisis emosi dan topik ini memungkinkan pemahaman yang lebih mendalam mengenai hubungan antara emosi masyarakat dan isu yang dibicarakan. Dengan pendekatan tersebut, proyek ini diharapkan dapat menghasilkan analisis yang lebih komprehensif, akurat, dan kontekstual mengenai respons publik terhadap OTT Abdul Wahid.

1.2 Tujuan

Tujuan dari proyek ini adalah sebagai berikut:

1. Menganalisis emosi publik terhadap isu Operasi Tangkap Tangan (OTT) Gubernur Riau Abdul Wahid berdasarkan komentar di platform *YouTube*.

2. Mengidentifikasi topik atau pihak yang menjadi sasaran emosi publik (seperti Gubernur, KPK, atau pemerintah) menggunakan metode *Latent Dirichlet Allocation* (LDA).
3. Melakukan labeling emosi menggunakan dua metode, yaitu IndoBERT dan EmoLex, serta membandingkan akurasi dan konsistensi hasilnya.
4. Menghubungkan hasil analisis emosi dan topik untuk memahami konteks kemunculan emosi tertentu pada isu yang dibahas.
5. Menyajikan hasil analisis dalam bentuk visualisasi data yang dapat digunakan sebagai dasar evaluasi respon publik dan pertimbangan kebijakan bagi lembaga atau pihak terkait.

1.3 Manfaat

Manfaat dari proyek ini adalah sebagai berikut:

1. Bagi pemerintah dan lembaga penegak hukum, dapat memberikan gambaran yang lebih terukur mengenai opini, sentimen, dan emosi masyarakat terhadap tindakan hukum yang dilakukan, sehingga dapat digunakan sebagai bahan evaluasi kebijakan publik dan strategi komunikasi.
2. Bagi masyarakat umum, dapat memberikan pemahaman yang lebih jelas tentang bagaimana opini dan emosi publik terbentuk dan berkembang di media sosial, terutama dalam isu-isu penegakan hukum.
3. Bagi pengembangan teknologi NLP Bahasa Indonesia, perbandingan IndoBERT dan EmoLex dapat memberikan masukan mengenai efektivitas metode labeling emosi, serta mendukung pengembangan model yang lebih baik untuk analisis opini publik kedepannya.

BAB 2
TINJAUAN PUSTAKA

2.1 Landasan Teori

1. Model IndoBERT

IndoBERT merupakan turunan dari *Bidirectional Encoder Representations from Transformers* (BERT) yang secara khusus dikembangkan untuk bahasa Indonesia. Model ini menerapkan arsitektur *transformer* dengan mekanisme *self-attention*, yang memungkinkan pemodelan konteks secara dua arah baik dari kiri ke kanan maupun sebaliknya. Kemampuan ini membuat IndoBERT efektif dalam memahami struktur kalimat, relasi antarkata, serta variasi ekspresi linguistik yang umum digunakan dalam komunikasi digital masyarakat Indonesia.

Menurut Koto, Rahimi, Lau, dan Baldwin (2020), IndoBERT menunjukkan performa unggul pada berbagai tugas pemrosesan bahasa alami untuk bahasa Indonesia, termasuk klasifikasi, ekstraksi informasi, dan analisis semantik. Keunggulan ini terutama disebabkan oleh pre-training yang memanfaatkan korpus berskala besar dalam bahasa Indonesia, sehingga model mampu menangkap pola dan karakteristik bahasa yang lebih representatif dibandingkan metode berbasis word embedding konvensional.

Dalam penelitian ini, IndoBERT digunakan untuk melakukan klasifikasi emosi pada komen tar YouTube. Pemilihan IndoBERT didasarkan pada kemampuannya dalam mengenali konteks dan nuansa bahasa Indonesia yang tidak baku, informatif, dan sering kali bersifat percakapan, seperti yang lazim ditemukan pada komentar media sosial

2. *Latent Dirichlet Allocation* (LDA)

Latent Dirichlet Allocation (LDA) merupakan metode *topic modeling* berbasis probabilistik yang mengasumsikan bahwa setiap dokumen tersusun atas sekumpulan topik, sedangkan setiap topik terbentuk dari distribusi kata tertentu. Dengan pendekatan generatif ini, LDA mampu mengidentifikasi struktur tematik tersembunyi dalam kumpulan teks tanpa memerlukan label atau supervisi. Sebagaimana dijelaskan oleh Vayansky dan Kumar (2020), LDA tetap menjadi salah satu metode *topic modeling* yang paling banyak digunakan karena interpretasinya yang jelas serta stabilitas modelnya dalam mengelompokkan dokumen berdasarkan kesamaan makna. Meskipun beberapa metode modern berbasis neural telah diperkenalkan, LDA tetap relevan dalam penelitian sosial dan analisis opini publik karena

mampu menghasilkan topik-topik yang mudah dipahami dan ditelusuri oleh peneliti. Pada teks pendek seperti komentar YouTube, LDA tetap dapat bekerja dengan baik apabila didukung dengan teknik prapemrosesan yang memadai, seperti pembentukan bigram atau penyaringan kata.

Dalam proyek ini, LDA digunakan untuk mengetahui topik-topik utama dalam komentar publik terkait kasus OTT Gubernur Riau. Informasi mengenai topik kemudian diintegrasikan dengan hasil klasifikasi emosi IndoBERT, sehingga penelitian dapat mengidentifikasi tidak hanya emosi yang muncul, tetapi juga objek atau isu yang menjadi sumber emosi tersebut.

3. NRC Emotion Lexicon (EmoLex)

NRC Emotion Lexicon (EmoLex) merupakan salah satu leksikon emosi yang banyak digunakan sebagai pendekatan berbasis kamus dalam analisis sentimen dan emosi. Dalam penelitian Wong dan Zhao (2023), EmoLex direpresentasikan sebagai metode leksikon yang bekerja dengan mencocokkan kata dalam teks dengan daftar kata yang telah diberi label emosi. Pendekatan ini dinilai efisien karena tidak membutuhkan pelatihan model dan mampu memberikan hasil yang konsisten berdasarkan aturan leksikal yang telah ditetapkan.

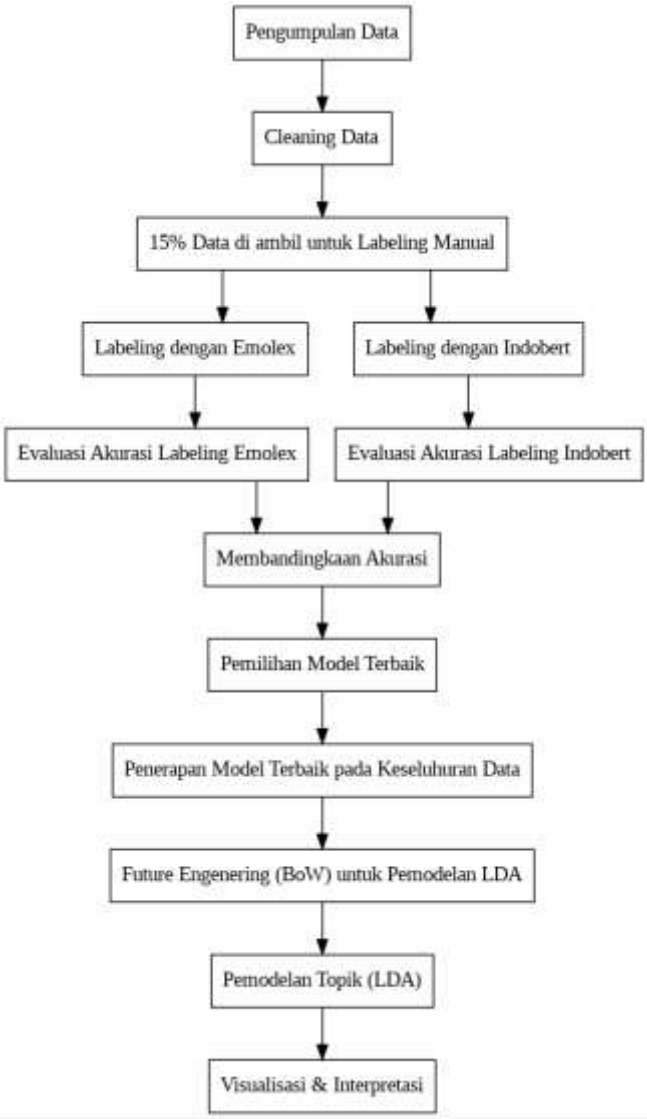
Wong dan Zhao (2023) juga menjelaskan bahwa metode leksikon seperti EmoLex memiliki keunggulan dalam hal transparansi dan kemudahan interpretasi, sehingga sering digunakan sebagai baseline dalam penelitian analisis emosi. Namun, pendekatan ini memiliki keterbatasan dalam menangani konteks yang kompleks, ironi, negasi, atau penggunaan bahasa informal yang umum ditemukan dalam komentar media sosial. Keterbatasan ini menyebabkan performanya cenderung lebih rendah dibandingkan model berbasis transformer.

Dalam proyek ini, EmoLex digunakan untuk menghasilkan pelabelan emosi otomatis pada komentar YouTube, yang kemudian dibandingkan dengan hasil klasifikasi IndoBERT serta divalidasi menggunakan label manual. Perbandingan ini dilakukan untuk menilai sejauh mana perbedaan pendekatan (leksikon vs. transformer(IndoBERT) memengaruhi kualitas hasil klasifikasi emosi pada teks tidak baku berbahasa Indonesia.

BAB 3
METODOLOGI

3.1 Diagram Alur dan Tahapan Analisis

Proses analisis emosi publik terhadap kasus OTT KPK terhadap Gubernur Riau dilakukan melalui beberapa tahapan yang saling berkaitan dan terstruktur secara sistematis. Tahapan-tahapan ini digambarkan secara visual dalam diagram jalur pada Gambar 3.1 yang menunjukkan alur penelitian mulai dari pengumpulan data hingga tahap interpretasi hasil. Melalui alur ini, penelitian berupaya mengubah data mentah dari media sosial menjadi informasi yang dapat menggambarkan reaksi emosional masyarakat serta topik utama yang diperbincangkan.



Gambar 3.1

Berdasarkan Gambar 3.1, tahapan proses analisis secara garis besar meliputi beberapa langkah berikut:

1. Pengumpulan Data

   Tahapan awal dilakukan dengan mengumpulkan data komentar dari platform YouTube menggunakan metode *web scraping*. Video yang dijadikan sumber dipilih berdasarkan relevansinya dengan kasus OTT KPK terhadap Gubernur Riau. Hasil scraping berupa teks komentar publik yang menjadi bahan utama untuk dianalisis lebih lanjut.

2. *Cleaning Data*

   Proses pembersihan teks bertujuan menghilangkan elemen-elemen yang tidak diperlukan dan dapat menyebabkan *noise* dalam analisis. Tahapan ini secara spesifik mencakup penghapusan URL, tanda baca, angka, emoji, serta elemen visual dan karakter lain yang tidak relevan. Normalisasi teks dilakukan untuk menyeragamkan format komentar. Langkah ini melibatkan perbaikan kata-kata yang tidak baku atau tidak formal, perluasan singkatan-singkatan yang lazim digunakan di media sosial, dan penanganan kata-kata asing yang tercampur, sehingga seluruh teks memiliki format yang konsisten dan siap diproses secara efektif oleh model analisis.

3. 15% Data Diambil untuk Labeling Manual

   Setelah seluruh data komentar berhasil dikumpulkan, tahap selanjutnya adalah melakukan pelabelan emosi pada setiap komentar, karena hasil pelabelan ini menjadi dasar bagi analisis emosi pada tahap-tahap berikutnya. Proses pelabelan diawali dengan membuat label manual pada sekitar 15% dari total data. Pada tahap ini, kami membaca setiap komentar dan menentukan kategori emosinya marah, jijik, takut, senang, sedih, netral, terkejut, dan percaya sehingga menghasilkan *ground truth* yang dijadikan acuan utama dalam mengevaluasi kualitas metode labeling.

4. Labeling dengan EmoLex dan Labeling dengan IndoBERT

   Dua metode labeling diterapkan pada data yang sama, yaitu EmoLex dan IndoBERT, dengan tujuan membandingkan akurasi keduanya terhadap label manual.

   - Pada pendekatan EmoLex, Setiap komentar dipecah menjadi token kata. Token-token ini kemudian dicocokkan dengan NRC Emotion Lexicon yang telah diterjemahkan ke Bahasa Indonesia. Emosi suatu komentar ditentukan

berdasarkan agregasi (penjumlahan skor) kategori emosi dari kata-kata yang ditemukan dalam kamus tersebut.

- Sedangkan pada pendekatan IndoBERT memanfaatkan model *transformer* yang canggih yang mampu memahami representasi konteks kalimat dan hubungan semantik antar kata. Komentar dimasukkan ke dalam model, yang melakukan tokenisasi *subword* dan mengubahnya menjadi *embeddings*. Berdasarkan pemahaman mendalam tentang struktur dan makna kalimat secara keseluruhan, IndoBERT kemudian menghasilkan prediksi emosi

5. Evaluasi Akurasi Labeling EmoLex dan Evaluasi Akurasi Labeling IndoBERT

   Setelah kedua hasil labeling otomatis diperoleh, keduanya dievaluasi dengan membandingkannya terhadap label manual.

6. Membandingkan Akurasi

   Hasil evaluasi menunjukkan bahwa IndoBERT memiliki akurasi lebih tinggi, sedangkan EmoLex lebih rendah karena kurang sensitif terhadap konteks, ironi, atau penggunaan bahasa yang tidak langsung.

7. Pemilihan Model Terbaik dan Penerapan Model Terbaik pada Keseluruhan Data

   Berdasarkan perbandingan akurasi, IndoBERT dipilih sebagai metode labeling utama dan digunakan untuk memberikan label emosi pada seluruh dataset, sehingga setiap komentar memiliki kategori emosi yang konsisten dan akurat sebelum masuk ke tahap berikutnya.

8. *Feature Engineering*

   Tahapan ini berfungsi untuk mengubah teks yang telah melewati proses labeling menjadi unit-unit kata individual atau biasa di sebut token. Cara ini paling cocok untuk LDA karena modelnya memang bekerja paling baik dengan hitungan kata yang sederhana, hasil topik jadi lebih stabil, mudah dipahami, dan cocok untuk komentar yang biasanya pendek.

9. Pemodelan Topik (LDA)

   Pemodelan topik menggunakan metode *Latent Dirichlet Allocation (LDA)*. Metode ini digunakan untuk mengidentifikasi dan mengelompokkan topik utama yang muncul dalam komentar publik, misalnya topik tentang kepercayaan terhadap KPK, opini terhadap pejabat daerah, atau tanggapan terhadap pemberantasan korupsi.

10. Visualisasi dan Interpretasi Hasil

Tahap terakhir adalah menyajikan temuan analisis dalam bentuk visual untuk memudahkan pemahaman pola dan kecenderungan yang muncul pada komentar publik. Visualisasi yang digunakan yaitu, wordcloud topik tiap emosi, tabel hubungan antara topik dan intensitas emosi tertentu, visualisasi. Melalui visualisi ini, dapat menginterpretasikan bagaimana publik merespons kasus OTT KPK terhadap Gubernur Riau, emosi apa yang paling dominan, serta isu atau topik apa yang paling banyak diperbincangkan. Hasil interpretasi ini kemudian menjadi dasar dalam menarik kesimpulan mengenai persepsi dan dinamika emosi masyarakat.

BAB 4
HASIL DAN PEMBAHASAN

4.1 Data Penelitian

Data yang digunakan dalam penelitian ini berupa komentar publik yang diperoleh dari 11 video YouTube yang membahas kasus OTT Gubernur Riau. Pengambilan data dilakukan melalui proses *scraping* menggunakan YouTube Data API sehingga seluruh data yang diperoleh merupakan komentar asli dari pengguna YouTube pada setiap video yang dipilih. Setiap komentar yang berhasil diambil terdiri dari dua komponen utama yaitu:

- *Author* (nama pengguna) : Identitas akun yang memberikan komentar
- *Comment* (isi komentar) : Teks komentar yang menjadi objek analisis penelitian

Data yang terkumpul kemudian disimpan dalam bentuk tabel *DataFrame* agar dapat diolah dan dianalisis secara terstruktur.

Setelah proses pengumpulan selesai, seluruh komentar menjalani tahap preprocessing untuk meningkatkan kualitas data sebelum masuk ke tahap analisis. Tahapan preprocessing yang dilakukan meliputi:

- pembersihan teks dasar (penghapusan URL, emoji, tanda baca, angka, dan karakter tidak relevan)
- penyeragaman format teks (lowercasing)
- normalisasi kata dan frasa tidak baku
- penghapusan huruf berulang
- penghapusan stopword tambahan yang dianggap tidak memberikan makna penting terhadap analisis
- penghapusan komentar yang identik (duplikasi)

Tabel 4.1.1 Data sebelum dilakukan *preprocessing*

No	Author	Comment
1	@pecis5767	Yg merugikan negara sampai trliunan kok gak di tangkap,malah pemimpin pro rakyat di bungkam,dulu bupati Meranti sekarang gebenur,
2	@kenken-q7n	wahid wahid mau nyari balek modal kau ya.riau ladang basah buat para koruptor
3	@Randiazuro.89	Dari cleaning service jadi gubernur 😂😂😂
4	@AsyariIbrahim-j8s	Kepada seluruh kepala daerah . Jika anggaran kita udah dikurangi . Jangan dipaksakan membangun yg kita tidak mampu . Berbuat saja sesuai kemampuan . Kerjanya yg lurus - lurus aja .

…	…	…
2574	@RealmeNabila-i2g	Riau oh Riau kenapa dulu kenapa sekarang ...😂

Tabel 4.1.2 Data setelah dilakukan *preprocessing*

No	Author	clean_comment
1	@pecis5767	yang merugikan negara sampai triliun tidak di tangkap malah pemimpin pro rakyat di bungkam dulu bupati meranti sekarang gebenur
2	@kenken-q7n	wahid wahid mau nyari balek modal kau yariau ladang basah buat para koruptor
3	@Randiazuro.89	dari pelayan kebersihan jadi gubernur
4	@AsyariIbrahim-j8s	kepada seluruh kepala daerah jika anggaran kita udah dikurangi jangan dipaksakan membangun yang kita tidak mampu berbuat saja sesuai kemampuan kerja kerja yang lurus lurus saja
…	…	…
2199	@RealmeNabila-i2g	riau oh riau kenapa dulu kenapa sekarang

Berdasarkan 2 tabel diatas, jumlah komentar mengalami penyusutan dari 2.575 komentar menjadi 2.199 komentar. Pengurangan ini terjadi karena beberapa komentar teridentifikasi sebagai duplikasi, ada komentar yang hilang seluruh isinya setelah dibersihkan, atau berubah menjadi teks yang tidak lagi bermakna sehingga tidak layak digunakan dalam analisis lanjutan.

4.2 **Evaluasi Model Labeling Emosi**

Pada tahap ini dilakukan evaluasi terhadap dua metode pelabelan otomatis, yaitu EmoLex dan IndoBERT, dengan menggunakan label manual sebagai ground truth. Evaluasi difokuskan pada nilai akurasi, precision, recall, dan f1-score untuk menentukan metode yang paling tepat digunakan dalam proses pelabelan emosi pada keseluruhan dataset.

- Evaluasi Model EmoLex

Tabel 4.2.1 Evaluasi Emolex

Emosi	Precision	Recall	F1-Score	Support
Jijik	0.14	0.54	0.23	13
Marah	0.36	0.14	0.20	109
Netral	0.30	0.50	0.37	64
Percaya	0.15	0.28	0.19	46
Sedih	0.22	0.07	0.11	27
Senang	0.23	0.09	0.13	54
Takut	0.11	0.17	0.13	18
Terkejut	0.00	0.00	0.00	17

Accuracy	—	—	**0.22**	348
Macro Avg	0.19	0.22	0.17	348
Weighted Avg	0.25	0.22	0.20	348

Model berbasis leksikon EmoLex diuji menggunakan data berlabel manual dan menghasilkan akurasi 0.22. Nilai precision dan recall pada sebagian besar kelas sangat rendah, sebagaimana ditampilkan pada Tabel 4.2.1. Kinerja yang lemah ini menunjukkan bahwa pendekatan berbasis kamus kurang mampu menangkap konteks, ironi, atau struktur kalimat yang umum ditemukan pada komentar YouTube yang bersifat informal dan tidak baku.

- Evaluasi Model IndoBERT

Tabel 4.2.2 Evaluasi IndoBERT

Emosi	Precision	Recall	F1-Score	Support
Jijik	0.92	0.92	0.92	13
Marah	0.90	0.91	0.91	100
Netral	0.84	0.90	0.87	60
Percaya	0.90	0.88	0.89	43
Sedih	0.96	0.81	0.88	27
Senang	0.88	0.94	0.91	54
Takut	0.86	0.75	0.80	16
Terkejut	0.93	0.82	0.88	17
Accuracy	—	—	**0.89**	330
Macro Avg	0.90	0.87	0.88	330
Weighted Avg	0.89	0.89	0.89	330

Model IndoBERT menunjukkan performa yang jauh lebih baik dengan akurasi 0.89 dan metrik yang tinggi pada hampir seluruh kelas emosi (Tabel 4.2.2). Nilai precision, recall, dan f1-score yang stabil mengindikasikan bahwa IndoBERT mampu memahami konteks linguistik secara lebih mendalam, termasuk variasi bahasa informal yang sering muncul pada komentar media sosial.

- Interpretasi dan Pemilihan Model Terbaik

Hasil evaluasi menunjukkan perbedaan performa yang sangat signifikan antara kedua metode. EmoLex hanya mencapai akurasi 22%, sedangkan IndoBERT mencapai 89% dengan kualitas prediksi yang jauh lebih baik di semua kelas emosi. Oleh karena itu, IndoBERT dipilih

sebagai metode utama pelabelan emosi untuk keseluruhan dataset karena mampu menghasilkan label yang lebih akurat dan konsisten, sehingga mendukung analisis topik dan interpretasi penelitian pada tahap selanjutnya.

4.3 Analisis Emosi

Analisis emosi dilakukan dengan menerapkan model IndoBERT pada seluruh dataset komentar YouTube untuk mengidentifikasi kecenderungan emosi publik terhadap kasus OTT Gubernur Riau. Hasil analisis menunjukkan bahwa emosi marah menjadi emosi yang paling dominan, diikuti oleh emosi senang dan netral. Dominasi emosi marah mencerminkan kuatnya reaksi negatif masyarakat terhadap praktik korupsi, sementara kemunculan emosi senang dan percaya menunjukkan adanya apresiasi serta kepercayaan publik terhadap upaya penegakan hukum. Distribusi emosi secara lengkap disajikan pada Tabel 4.3.1.

Tabel 4.3.1 Distribusi emosi

Emosi	Jumlah
Marah	665
Senang	488
Netral	476
Percaya	314
Sedih	97
Jijik	57
Terkejut	55
Takut	47

Untuk memperdalam pemahaman terhadap ekspresi emosi tersebut, dilakukan visualisasi word cloud pada masing-masing kategori emosi. Word cloud ini menampilkan kata-kata yang paling sering muncul pada setiap emosi sehingga memberikan gambaran konteks dan fokus perhatian masyarakat.

Tabel 4.3.2 WordCloud

Berdasarkan hasil word cloud, emosi marah dan jijik didominasi oleh kata-kata seperti “korupsi”, “koruptor”, “gubernur”, “pejabat”, dan “rakyat”, yang menegaskan kemarahan serta rasa muak masyarakat terhadap penyalahgunaan jabatan. Pada emosi senang dan percaya, kata “KPK” muncul secara dominan, menunjukkan apresiasi dan kepercayaan publik terhadap lembaga penegak hukum. Sementara itu, emosi sedih, takut, dan terkejut menampilkan kata-kata yang berkaitan dengan kerugian negara, penetapan tersangka, serta kekhawatiran terhadap dampak kasus tersebut.

Secara keseluruhan, visualisasi distribusi emosi dan word cloud menunjukkan bahwa respons publik didominasi oleh emosi negatif, namun tetap disertai emosi positif terhadap proses penegakan hukum. Hasil ini menjadi dasar untuk analisis lanjutan mengenai keterkaitan antara topik dan emosi pada tahap berikutnya.

4.4 Feature Engineering

Setelah semua komentar sudah di labelling, Langkah berikutnya adalah Tokenisasi, dimana setiap unit kata dipecah menjadi unit-unit kata individual atau yang disebut token. Setelah pembentukan token, dilakukan penghapusan tanda baca(titik, koma, dan tanda seru) dari daftar token karena tidak memberikan kontribusi makna dari dokumen dan penghapusan stopword. Seluruh rangkaian operasi pembersihan dan tokenisasi ini kemudian diterapkan ke setiap baris data, menghasilkan kolom baru, yaitu tokens, yang berisi daftar kata-kata inti yang telah bersih untuk setiap komentar yang siap digunakan utuk LDA.

Tabel. 4.4.1 Tokenisasi Komentar

	Clean_comment	IndoBERT_Label	Tokens
1	yang merugikan negara sampai triliun tidak di tangkap malah pemimpin pro rakyat di bungkam dulu bupati meranti sekarang gebenur	Marah	['merugikan', 'negara', 'triliun', 'tangkap', 'pemimpin', 'pro', 'rakyat', 'bungkam', 'bupati', 'meranti', 'gebenur']
2	wahid wahid mau nyari balek modal kau yariau ladang basah buat para koruptor	Senang	['wahid', 'wahid', 'nyari', 'balek', 'modal', 'kau', 'yariau', 'ladang', 'basah', 'koruptor']
3	dari pelayan kebersihan jadi gubernur	Netral	['pelayan', 'kebersihan', 'gubernur']
4	kepada seluruh kepala daerah jika anggaran kita udah dikurangi jangan dipaksakan membangun yang kita tidak mampu berbuat saja sesuai kemampuan kerja kerja yang lurus lurus saja	Takut	['kepala', 'daerah', 'anggaran', 'udah', 'dikurangi', 'dipaksakan', 'membangun', 'berbuat', 'sesuai', 'kemampuan', 'kerja', 'kerja', 'lurus', 'lurus']
…	…	…	…
2199	riau oh riau kenapa dulu kenapa sekarang	Netral	[‘riau’, ‘oh’, ‘riau’]

4.5 Pemodelan LDA

Tahap selanjutnya adalah Penentuan jumlah topik ideal untuk pemodelan Latent Dirichlet Allocation (LDA). Proses ini dimulai dengan eksplorasi rentang topik yang luas, di mana model LDA dilatih berulang kali untuk jumlah topik yang bervariasi dari 5 hingga 100. Tujuan utama dari eksplorasi ini adalah menghitung Skor Koherensi (*Coherence Score*) untuk setiap model

yang dilatih. Skor Koherensi, yang mengukur seberapa sering kata-kata teratas dalam suatu topik muncul bersama-sama dalam dokumen, berfungsi sebagai metrik objektif untuk menilai kualitas dan kemudahan interpretasi topik yang dihasilkan secara statistik. Berdasarkan hasil perhitungan awal, didapatkan bahwa Skor Koherensi tertinggi secara absolut dicapai pada jumlah topik K=29 dengan Skor Koheren 0.4979329143395437.

Meskipun model dengan K=29 secara teknis memiliki koherensi statistik terbaik, hasil ini menimbulkan kesulitan signifikan dalam tahap interpretasi. Mengidentifikasi dan memberikan label yang bermakna pada 29 topik yang sangat spesifik dan sering kali tumpang tindih memerlukan sumber daya dan waktu yang besar, dalam jumlah topik ini juga di temukan banyaknya topik dan bobot kata yang sama. Oleh karena itu, dilakukan eksplorasi ulang pada rentang topik yang lebih terkonsentrasi dan realistis untuk tujuan analisis, yaitu antara 5 hingga 10.

Tabel 4.5.1 Jumlah topik & skor koheren

Num_topics	Coherence_score
5	0.337805
6	0.357537
7	0.379747
8	0.368949
9	0.396375
10	0.404544

Dalam rentang ini, didapatkan bahwa model dengan 10 Topik menghasilkan Skor Koherensi tertinggi 0.4045439227497928. Keputusan akhir kemudian diambil untuk menggunakan 10 Topik (K=10) sebagai jumlah topik yang optimal. Pilihan ini didasarkan pada prinsip keseimbangan antara Koherensi Statistik yang memadai dan Kemudahan Interpretasi (*Interpretability*) yang tinggi, memungkinkan analisis yang lebih ringkas, fokus, dan relevan.

Berdasarkan hasil keputusan untuk memilih 10 topik berdasarkan kemudahan interpretasi, langkah selanjutnya adalah menganalisis kata-kata kunci dari setiap topik untuk menentukan label atau tema topik utama. Ringkasan dari interpretasi topik, termasuk kata-kata kunci yang mendefinisikannya, disajikan dalam Tabel berikut:

Tabel. 4.5.3 Kata Kunci dan Label Topik LDA

Topik	Kata-kata Kunci	Label/Topik Utama
1	Gubernur, kpk, riau, preman, nya, maling, jatah, kepala, sumut, hukum.	Tokoh politik, hukum dan otoritas regional
2	Riau, korupsi, gubernur, pejabat, orang, hukum, Prabowo, ya, jalan, Indonesia.	Isu korupsi regional (Riau) dan keterlibatan politik
3	Maling, riau, uang, rakyat, nya, d, pusat, udah,kaya, korup.	Perampasan uang rakyat oleh koruptor regional.
4	Korupsi, kpk, gubernur, mati, tu, riau, korupsi, negara, takut, hukuman.	Desakan hukuman berat bagi koruptor
5	Omon, koruptor, korupsi, kpk, klau, Jokowi, mata, riau,berantas, senyap.	Opini krisis terhadap Upaya pemberantasan korupsi (KPK & Pemerintah)
6	Tangkap, korupsi, kpk, nya, pake, si, bpk, somad, pasal, an.	Tindakan penangkapan dan audit Lembaga hukum (KPK & BPK)
7	Korupsi, ya politik, hukuman,takut, hidup, manusia,negara, y, masyarakat.	Korupsi sebagai isu etika, politik, dan dampak sosial
8	Koruptor, mati, rakyat, nya, hukum, asset, pejabat, hukuman, dpr, Indonesia.	Tuntutan hukuman mati dan penyitaan asset pejabat
9	Korupsi,pejabta, negara, kpk, berani,daerah, lemah, bebas, kelas, habis.	Kelemahan penegakan hukum dan oejabat yang kebal hukum
10	Ott, korupsi, proyek, orang wahid, periksa, gubernurnya, whoosh, pemberantasan, contohnya.	Operasi tangkap tangan (OTT) dan proyek pemerintah

Pada tabel 4.5.2 ditemukan bahwa diskusi publik terfokus secara intensif pada isu Korupsi dan Penegakan Hukum di Indonesia. Proses interpretasi topik dilakukan dengan menganalisis kata-kata kunci teratas (probabilitas tertinggi) pada setiap topik, dan kemudian merangkumnya

menjadi label/topik utama yang koheren. Topik 1, 2, dan 3 berpusat pada dimensi regional, menyoroti Isu Korupsi Regional yang melibatkan Gubernur dan lembaga seperti KPK, khususnya terkait wilayah Riau dan Sumut, dengan nada yang mencerminkan kemarahan atas Perampasan Uang Rakyat oleh 'maling'.

Topik-topik berikutnya mengungkapkan tuntutan publik yang sangat keras terhadap sanksi hukum. Topik 4 dan 8 secara eksplisit mewakili Desakan Hukuman Berat, dengan kata kunci seperti "mati", "hukuman", "koruptor", dan "aset", yang mencerminkan aspirasi publik agar koruptor dihukum mati dan aset mereka disita. Sementara itu, Topik 5, 6, dan 9 mengkritisi institusi. Topik 5 mencerminkan Opini Kritis terhadap Upaya Pemberantasan Korupsi, melibatkan pemerintah dan KPK. Topik 9 menyoroti Kelemahan Penegakan Hukum terhadap pejabat negara yang dianggap "kebal" atau "bebas", sementara Topik 6 membahas Tindakan Penangkapan dan Audit spesifik oleh KPK dan BPK. Terakhir, Topik 10 mengacu pada kasus spesifik, yaitu Operasi Tangkap Tangan (OTT) dan Proyek Pemerintah, dan Topik 7 menempatkan korupsi sebagai Isu Etika, Politik, dan Dampak Sosial yang lebih luas terhadap negara dan masyarakat. Dengan demikian, model LDA berhasil memetakan lanskap narasi publik menjadi sepuluh tema yang jelas. Langkah logis selanjutnya adalah menganalisis distribusi emosi pada setiap topik, sehingga dapat diidentifikasi emosi dominan yang mendorong narasi yang terkait dengan isu seputar korupsi dan penegakan hukum.

Tahap selanjutnya adalah menganalisis persentase topik pada setiap komentar. Langkah ini penting karena satu komentar bisa membahas beberapa isu sekaligus. Dengan melihat topik yang paling dominan, kita dapat mengetahui fokus utama tiap komentar dan seberapa kuat pembahasannya terhadap isu tertentu, seperti korupsi, hukuman, atau kinerja lembaga. Hasil analisis ini juga menjadi dasar untuk melihat emosi utama yang muncul pada setiap topik, sehingga kita bisa memahami bagaimana perasaan publik terhadap isu-isu tersebut.

Tabel 4.5.2 Presentase Topik di Setiap Komentar

Comment	Presentase Bobot Topik
Comment 0	[(7, np.float32(0.6090471)), (8, np.float32(0.21446694)), (9, np.float32(0.13277349))]
Comment 1	[(7, np.float32(0.6590644)), (9, np.float32(0.28489876))]

Comment 2	[(0, np.float32(0.29031962)), (1, np.float32(0.021222249)), (2, np.float32(0.01931302)), (3, np.float32(0.01981556)), (4, np.float32(0.5512977)), (5, np.float32(0.017166842)), (6, np.float32(0.017030593)), (7, np.float32(0.023767972)), (8, np.float32(0.023753196)), (9, np.float32(0.0163132))]
Comment 3	[(0, np.float32(0.2611275)), (2, np.float32(0.5593503)), (4, np.float32(0.1441321))]
Comment 4	[(0, np.float32(0.011459985)), (1, np.float32(0.4035535)), (2, np.float32(0.19034494)), (7, np.float32(0.011499871)), (8, np.float32(0.011491946)), (9, np.float32(0.33655518))]
…..	……
Comment 2198	[(0, np.float32(0.8231809)), (1, np.float32(0.021211412)), (2, np.float32(0.019303383)), (3, np.float32(0.019805482)), (4, np.float32(0.018516714)), (5, np.float32(0.017158132)), (6, np.float32(0.017021943)), (7, np.float32(0.023755925)), (8, np.float32(0.02374119)), (9, np.float32(0.016304927))]

Untuk menentukan sebuah komentar termasuk ke topik tertentu, setiap komentar dianalisis berdasarkan bobot yang dihasilkan oleh model LDA. Bobot terbesar menunjukkan topik yang paling dominan, sehingga topik dengan nilai tertinggi dianggap sebagai topik utama komentar tersebut. Sebagai contoh, pada tabel 4.5.2 comment 0, model LDA menghasilkan beberapa bobot topik, dan Topik 7 memiliki bobot paling besar yaitu 0.60 dibandingkan topik lainnya. Karena topik ini yang paling dominan, komentar tersebut kemudian dikategorikan ke dalam Topik 7, yaitu topik yang membahas korupsi sebagai isu etika, politik, dan dampak sosial. Hal ini bisa diterapkan dengan cara yang sama pada semua komentar, sehingga setiap komentar dapat dipetakan ke dalam topik yang paling sesuai berdasarkan bobot dominannya dan kemudian disesuaikan dengan label topik yang telah ditetapkan sebelumnya. Dengan demikian, seluruh komentar dalam dataset dapat diklasifikasikan dengan runtut ke dalam topik-topik utama yang berhasil diidentifikasi oleh model.

4.6 Analisis Topik dan Emosi

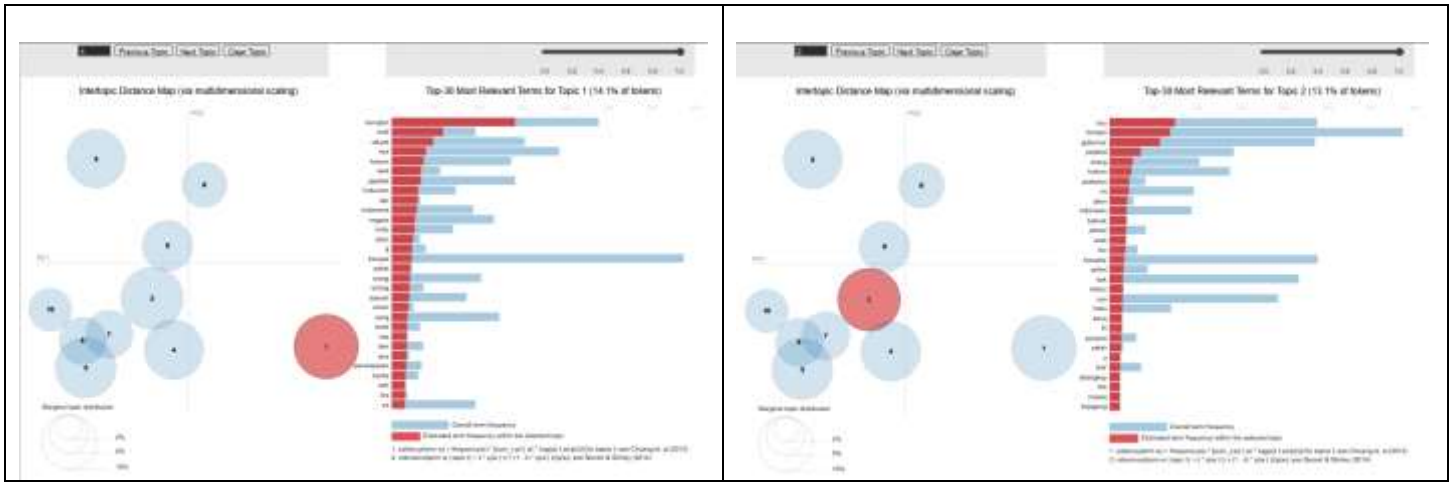
Analisis Topik dan Emosi dilakukan untuk mengidentifikasi sentimen publik yang menyertai 10 topik bahasan yang teridentifikasi melalui Latent Dirichlet Allocation (LDA), memberikan wawasan komprehensif mengenai perspektif pengguna terhadap isu-isu yang dibahas.

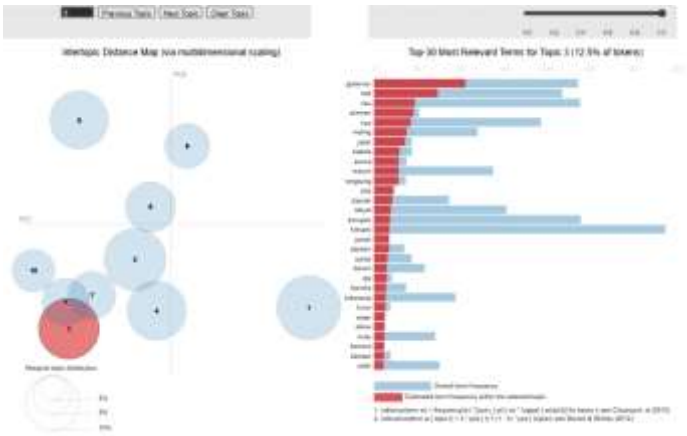
Tabel 4.6.1 Distribusi Emosi pada Setiap Topik

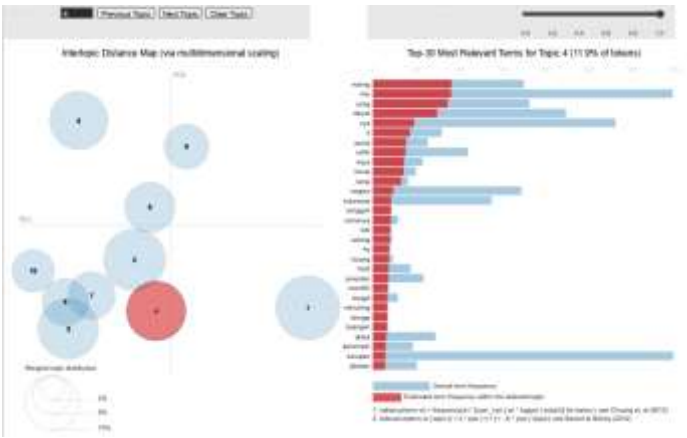
Topik	Jijik	Marah	Netral	Percaya	Sedih	Senang	Takut	Terkejut
1	10	69	91	47	11	72	12	10
2	16	92	42	50	15	58	8	3
3	5	80	40	42	16	57	5	2
4	3	57	39	23	8	54	3	5
5	1	56	47	23	5	41	3	3
6	3	61	49	21	10	29	1	6
7	1	53	36	19	3	37	1	2
8	8	87	49	45	14	54	6	7
9	7	66	52	31	9	57	2	9
10	3	44	31	13	6	29	6	8

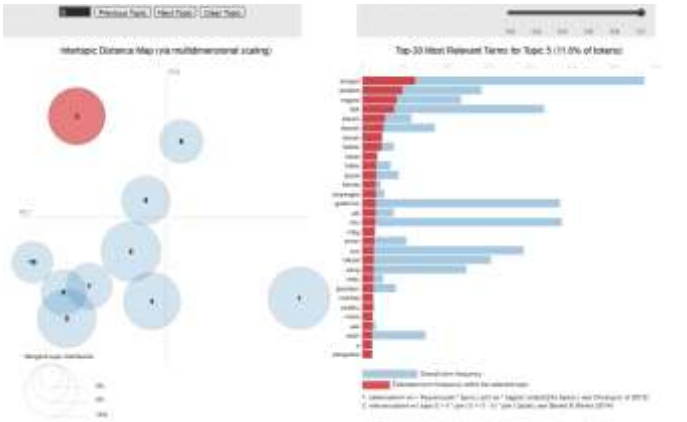
Berdasarkan Tabel Tabel 4.6.2 menunjukkan bahwa emosi Marah mendominasi di hampir semua topik dengan frekuensi tertinggi pada Topik 2 (92) dan Topik 8 (87), diikuti oleh emosi Netral dan Senang yang signifikan. Dominasi kemarahan ini merefleksikan kuatnya frustrasi dan ketidakpuasan publik terhadap isu-isu yang dibahas. Meskipun demikian, emosi Netral (tertinggi di Topik 1 dengan 91) mengindikasikan adanya porsi diskusi yang lebih faktual, sementara tingginya emosi Senang (tertinggi di Topik 1 dengan 72) dan Percaya (tertinggi di Topik 2 dengan 50) menunjukkan adanya apresiasi atau optimisme publik terhadap upaya penindakan hukum atau potensi hasil yang positif. Topik 1 merupakan titik fokus utama diskusi yang paling kompleks.

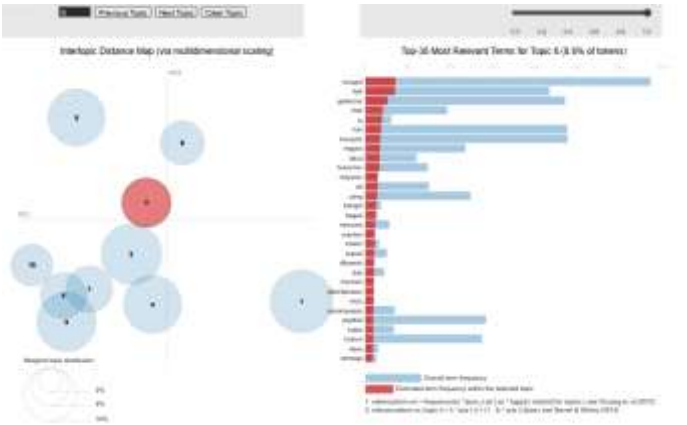
Tabel 4.6.2 Visualisasi LDA per Topik

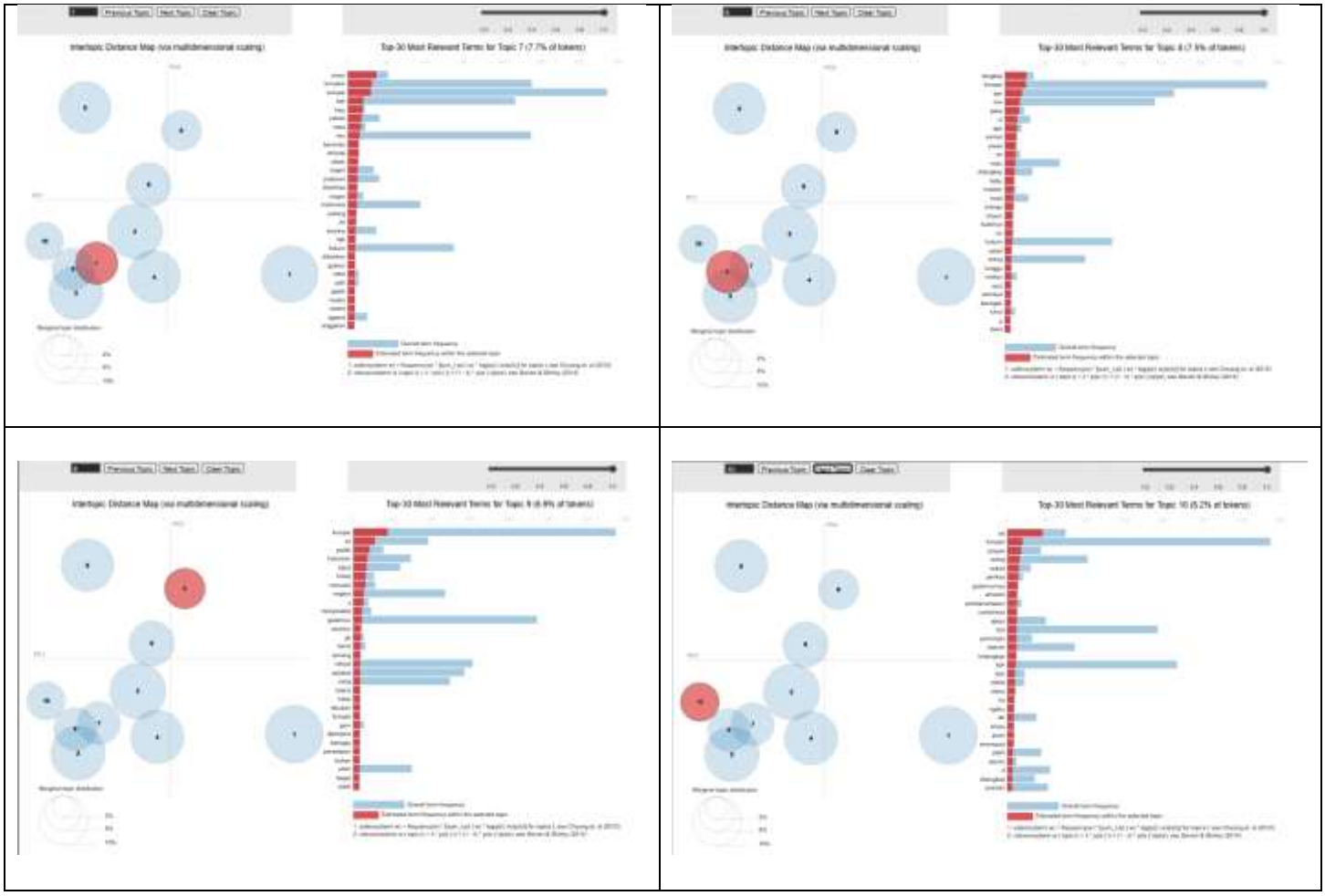



Previous Topic
Next Topic
Clear Topic
Interactive Distance Map (via multidimensional scaling)
Top-30 Most Relevant Terms for Topic 3 (12.5% of tokens)


Previous Topic
Next Topic
Clear Topic
Interactive Distance Map (via multidimensional scaling)
Top-30 Most Relevant Terms for Topic 4 (11.9% of tokens)


Previous Topic
Next Topic
Clear Topic
Interactive Distance Map (via multidimensional scaling)
Top-30 Most Relevant Terms for Topic 5 (11.0% of tokens)


Previous Topic
Next Topic
Clear Topic
Interactive Distance Map (via multidimensional scaling)
Top-30 Most Relevant Terms for Topic 6 (9.9% of tokens)



Berdasarkan Tabel 4.6.2 yang merupakan interpretasi visualisasi LDA lebih lanjut mengonfirmasi bahwa semua topik, yang diwakili oleh gelembung-gelembung pada peta dua dimensi, berpusat pada isu korupsi, penindakan hukum oleh KPK (OTT), dan peran pejabat/gubernur. Kata-kata kunci yang relevan seperti *koruptor, ott, gubernur, maling, kpk, riau, pejabat, hukuman, uang,* dan *negara* menjadi fokus utama pembahasan. Tingginya frekuensi emosi Marah, terutama pada Topik 1 dan 8, sangat mungkin dipicu oleh kata kunci yang mengandung unsur hukuman dan tindak kriminal seperti *maling*, *koruptor*, *mati*, dan *perampasan harta*, yang mencerminkan tuntutan publik akan penegakan hukum yang tegas. Di sisi lain, kedekatan dan tumpang tindih antar gelembung topik menunjukkan bahwa topik-topik tersebut saling terkait erat, mengindikasikan bahwa diskusi mengenai korupsi di berbagai daerah atau pada tingkatan pejabat yang berbeda masih menggunakan kosakata dan memicu sentimen yang serupa.

BAB 5
KESIMPULAN

Hasil proyek ini menunjukkan bahwa respon masyarakat terhadap kasus Operasi Tangkap Tangan (OTT) Gubernur Riau Abdul Wahid didominasi oleh emosi marah. Emosi ini terutama muncul pada topik-topik yang berkaitan dengan isu korupsi daerah serta tuntutan pemberian hukuman yang lebih berat bagi pejabat publik yang terlibat. Dominasi emosi marah menunjukkan tingginya ketidakpuasan dan kekecewaan publik terhadap korupsi yang masih sering terjadi, serta rendahnya kepercayaan terhadap integritas sebagian pejabat dan sistem pemerintahan di Indonesia.

Selain emosi marah, emosi netral juga ditemukan dalam jumlah yang cukup signifikan, khususnya pada komentar yang berfokus pada penyampaian informasi, opini faktual, serta pembahasan peran tokoh politik dan lembaga hukum. Hal ini menunjukkan bahwa tidak semua respons publik bersifat emosional, tetapi juga juga mencerminkan sikap rasional dan analitis dalam menanggapi kasus OTT. Sementara itu, emosi positif seperti senang dan percaya muncul pada topik yang membahas tindakan KPK, hal ini menandakan adanya apresiasi, optimisme, dan dukungan masyarakat terhadap upaya penegakan hukum yang dilakukan melalui OTT.

Secara keseluruhan, proyek ini menunjukkan bahwa respon publik terhadap isu OTT bersifat kompleks. Masyarakat tidak hanya mengekspresikan kemarahan terhadap pelaku korupsi dan sistem hukum yang dianggap lemah, tetapi juga menunjukkan dukungan serta harapan terhadap lembaga penegak hukum. Hasil analisis ini dapat dimanfaatkan sebagai bahan evaluasi bagi pemerintah dan lembaga terkait dalam memahami persepsi publik, merancang strategi komunikasi yang lebih efektif, serta memperkuat kepercayaan masyarakat melalui penegakan hukum yang konsisten, tegas, dan transparan.

DAFTAR PUSTAKA

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). *IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP*. [2011.00677] IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP

Vayansky, I., & Kumar, S. (2020). *A review of topic modeling methods*. Information Systems, 94, 101582. (PDF) A review of topic modeling methods

Wong, T., & Zhao, W. (2023). *Automated measures of sentiment via transformer- and lexicon-based sentiment analysis*. (PDF) Automated Measures of Sentiment via Transformer- and Lexicon-Based Sentiment Analysis (TLSA)

Ogunleye, B., Lancho Barrantes, B. S., & Zakariyyah, K. I. (2025). *Topic modelling through the bibliometrics lens and its technique. Artificial Intelligence Review*. https://www.researchgate.net/publication/387749558_Topic_modelling_through_the_bibliometrics_lens_and_its_technique

KONTRIBUSI KELOMPOK

Nama Anggota	**Kontribusi**			
	Proposal	Progres 1	Progres 2	Laporan Akhir
1. Muhammad Fahri Novrian (1314623014)	-BAB 3 Rancangan Proyek -Kesimpulan	-Penyelesaian code sampai dengan tahap proprocesing	-Penyelesaian code sampai dengan tahap Pemodelan LDA	-Menambahkan code visualisasi wordcloude -BAB 4 Hasil dan Pembahasan
2. Christine Aprilia Putri (240311554046)	-BAB 3 Rancangan Proyek -Kesimpulan	-Pemaparan progres 1	-Pemaparan progres 2 -Kesimpulan	-Mencari Model Training Yang optimal -BAB 4 Hasil dan Pembahasan
3. Ayu Wulan Anggraeni Putri (24031554177)	-BAB 2 Tinjauan Pustaka			
4. Fridania Nisa Calita (24031554208)	-BAB 1 Pendahuluan -Kesimpulan	-Pemaparan progres 1	-Pemaparan progres 2 -Kesimpulan	-Mencari model training yang optimal -Menambahkan Code dengan Coherence Score tertinggi

				-BAB 4 Hasil dan Pembahasan -Kesimpulan

Link GirHub: https://github.com/christine2646/UAS-KELOMPOK-2-PEMROSESAN-TEKS-ANALISIS-EMOSI-PUBLIK.git