

# Supervised Learning untuk Analisis Sentimen Tweet Pilpres 2024:

Klasifikasi Sentiment, Topik, Kandidat dengan GPT-4 API dan XLM-RoBERTa

Oleh:

Reyner Ongkowijoyo

Muhammad Fath Rajihan Nafie

Billy Jonathan



**SEMIFINAL  
BIG DATA CHALLENGE  
2024**



**Puspresnas**  
Pusat Prestasi Nasional

**BPTI**  
Badan Pengembangan Talenta Indonesia

**Telkom**  
University

**Kampus**  
**Merdeka**  
INDONESIA JAYA

# **Supervised Learning untuk Analisis Sentimen Tweet Pilpres 2024:**

## **Klasifikasi Sentiment, Topik, Kandidat dengan GPT-4 API dan XLM-RoBERTa**

### **ABSTRAK**

Pada tahun 2024 menjadi tahun politik bagi Indonesia, hal ini dikarenakan pada tahun tersebut diselenggarakannya pemilihan umum secara serentak berdasarkan keputusan KPU RI No 21 Tahun 2022. Di era politik modern, pendekatan melalui teknologi merupakan strategi yang sangat tepat untuk para koalisi pendukung calon presiden dan wakil presiden untuk menyebarluaskan visi, misi, program unggulan, dan citra diri. Penelitian ini bertujuan mengidentifikasi sentimen, topik, dan kandidat dalam tweet selama kampanye pemilihan presiden Indonesia 2024 menggunakan pendekatan supervised learning dan model XLM-RoBERTa. Sebuah pendekatan unik diajukan dengan menggunakan GPT-4o melalui API untuk proses pelabelan, yang mencapai akurasi lebih dari 90% dalam mengidentifikasi kandidat, sentimen, dan topik. Setelah validasi, sampel data diambil untuk fine-tuning model XLM-RoBERTa sebagai model utama dalam klasifikasi melalui data yang sudah dilabeli GPT-4o. Hasil penelitian menunjukkan performa terbaik dengan akurasi sentimen 0.838516, akurasi topik 0.664720, dan F1-Score kandidat 0.864740. Analisis hasil klasifikasi model menunjukkan bahwa topik yang paling banyak dibahas adalah seputar dinamika pendukung dan atribut kandidat, bukan kebijakan politik. Dengan ini, kami harapkan setiap kandidat memberikan prioritas utama dalam komunikasi dan peningkatan citra diri.

Kata kunci: Pilpres 2024; Supervised Learning; Sentiment Analysis; XLM-RoBERTa; GPT-4o API.

### **I. PENDAHULUAN**

#### **1.1 Latar Belakang**

Pada tahun 2024 menjadi tahun politik bagi Indonesia, hal ini dikarenakan pada tahun tersebut diselenggarakannya pemilihan umum secara serentak berdasarkan keputusan KPU RI No 21 Tahun 2022 (KPU RI, 2022), yang salah satunya termasuk juga pemilihan presiden dan wakil presiden. Pemilihan Umum adalah sebuah kompetisi untuk mengisi jabatan-jabatan politik di pemerintahan yang didasarkan pada pilihan formal dari warga negara yang memenuhi syarat (Astrianti Defretes, 2023). Pemilihan presiden dan wakil presiden RI pada tahun 2024 ini terdapat tiga pasangan calon presiden dan wakil presiden yang bersaing mendapatkan dukungan dari seluruh rakyat Indonesia terutama

pada masa kampanye. Kampanye merupakan serangkaian tindakan komunikasi yang terencana dengan tujuan untuk menciptakan efek tertentu pada sejumlah besar khalayak yang dilakukan secara berkelanjutan pada kurun waktu tertentu (Wibowo, 2023).

Dalam konteks kampanye di era politik modern, pendekatan melalui teknologi merupakan strategi yang sangat tepat untuk para koalisi pendukung calon presiden dan wakil presiden untuk menyebarluaskan visi, misi, program unggulan, dan citra diri, karena dengan teknologi, perkembangan informasi terhadap suatu isu-isu tertentu terjadi dengan sangat cepat, salah satunya melalui media sosial. Media sosial telah mengubah paradigma komunikasi politik di era modern karena media sosial telah menjadi sarana komunikasi yang kuat dan luas digunakan oleh individu, partai politik, dan kandidat dalam kampanye politik (Salman Farid, 2023).

Selama masa kampanye, aktivitas pengguna dalam platform X ini menjadi sangat padat. Setiap hari, jutaan pengguna platform X membuat, membagikan, dan merespons konten yang berkaitan dengan calon presiden dan wakil presiden, seperti visi, misi, latar belakang calon, dan program kerja. Hal ini menghasilkan besarnya data yang tersimpan dalam platform X dengan berbagai interaksi, seperti tweet, retweet, mention, like, komentar, dan lain lain.

Kumpulan data yang sangat besar dan kompleks biasa disebut dengan istilah *Bigdata*, dalam ilmu komputer, *big data* tidak dapat diolah dengan menggunakan perangkat lunak konvensional atau aplikasi pemrosesan data konvensional lainnya. Karena Teknologi *big data* memiliki karakteristik 3V yaitu, *Volume* (ukuran yang sangat besar), *Variety* (beragam macam data), dan *Velocity* (kecepatan akses data) (Maryanto, 2017).

Penelitian ini memanfaatkan pendekatan *supervised learning* pada media sosial X selama masa kampanye pemilihan presiden dan wakil presiden. *Supervised learning* adalah metode klasifikasi pada dataset yang diberikan label untuk mengklasifikasikan kelas yang tidak dikenal (Roihan et al., 2019). Dalam mengklasifikasikan data pada penelitian ini menggunakan model pra-terlatih yaitu XLM-RoBERTa dengan data *prelabeled* GPT-4o. Berdasarkan penelitian dari (Conneau et al., 2019), XLM-RoBERTa merupakan model bahasa multibahasa yang besar, yang dilatih pada data CommonCrawl yang difilter sebesar 2,5 TB yang memuat 100 bahasa dan GPT-4 adalah model kecerdasan buatan berbasis transformer yang dikembangkan OpenAI dan telah dilatih

dengan sejumlah besar data teks dari internet, karena hal inilah penelitian ini menawarkan peluang besar untuk mendapatkan wawasan mendalam tentang pandangan dan preferensi publik terhadap calon presiden dan wakil presiden dengan mengidentifikasi isu utama atau topik yang menjadi *highlight* para pengguna media sosial dan menganalisis sentimen atau emosi yang terkandung dalam opini atau teks.

### 1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, diperoleh rumusan masalah dalam penelitian ini sebagai berikut,

1. Bagaimana cara menyiapkan dari *big data* yang berasal dari media sosial X untuk membuat model klasifikasi sentimen, topik, dan kandidat?
2. Bagaimana cara mengidentifikasi sentimen, topik, dan kandidat pada *tweet* antar pengguna media sosial X selama masa kampanye pemilu 2024?
3. Apa saja informasi yang diperoleh dari hasil klasifikasi yang dilakukan model klasifikasi sentimen, topik, dan kandidat pada *tweet* antar pengguna media sosial X selama masa kampanye pemilu 2024?

### 1.3 Tujuan Penelitian

Berdasarkan rumusan masalah pada penelitian ini, maka tujuan penelitian ini sebagai berikut,

1. Mengetahui cara menyiapkan data besar yang berasal dari media sosial X untuk membuat model klasifikasi sentimen, topik, dan kandidat.
2. Mengidentifikasi sentimen, topik, dan kandidat yang terkandung pada *tweet* antar pengguna media sosial X selama masa kampanye pemilu 2024.
3. Mengetahui informasi yang diperoleh dari hasil klasifikasi yang dilakukan model klasifikasi sentimen, topik, dan kandidat pada *tweet* antar pengguna media sosial X selama masa kampanye pemilu 2024.

### 1.4 Manfaat

Berdasarkan tujuan penelitian pada penelitian ini, maka manfaat yang diperoleh dari penelitian ini sebagai berikut,

1. Bagi Civitas akademik

Penelitian ini bisa mengembangkan Ilmu Pengetahuan terutama pada bidang *data science*, *big data analytics* dan analisis media sosial dan menjadi

refrensi studi yang dapat digunakan oleh peneliti lain untuk membandingkan hasil, menguji hipotesis dan memperluas jangkauan penelitian yang akan datang.

## 2. Bagi Industri

Peneliti bidang industri dapat mengimplementasikan algoritma dan model yang digunakan dalam penelitian ini untuk penelitian industri atau mengembangkan bisnis yang dapat membantu merancang strategi bisnis yang lebih efektif

## 3. Bagi Masyarakat:

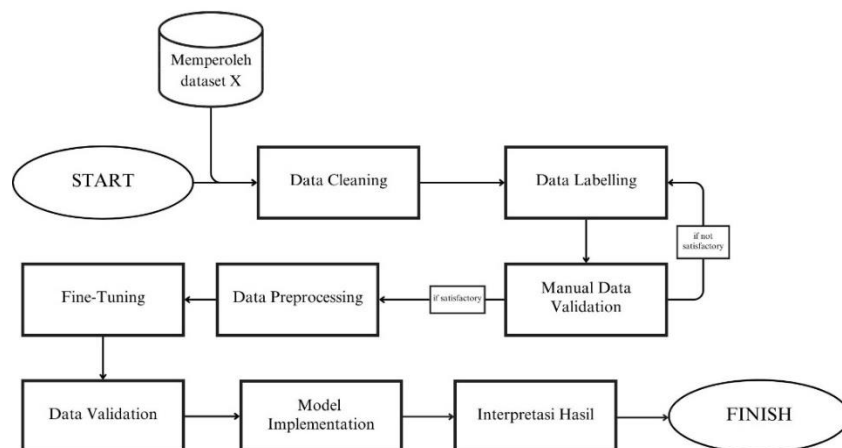
Penelitian ini dapat membantu meningkatkan kesadaran publik bahwa opini masyarakat dapat dianalisis dan digunakan untuk mempengaruhi kebijakan sosial.

## 4. Bagi Pembuat Kebijakan:

Penelitian ini dapat memberikan wawasan yang mendalam tentang isu-isu atau topik yang menjadi perhatian bagi masyarakat, sehingga pembuat kebijakan dapat membuat keputusan yang lebih tepat dan sesuai dengan kebutuhan masyarakat.

## II. METODOLOGI

Untuk mencapai tujuan penelitian ini, peneliti menggunakan *supervised learning* untuk mengenali sentiment teks, kandidat dibahas, dan topik yang dibahas. Hal ini dikarenakan *supervised learning* memiliki hasil prediksi yang lebih konsisten daripada *unsupervised learning* yang cenderung lebih sulit untuk diinterpretasikan. Oleh karena itu, metodologi yang akan digunakan pada penelitian ini akan mengikuti alur penelitian yang ditunjukkan pada gambar 1.



Gambar 1. Alur Penelitian

## 2.1 Dataset

Dataset yang digunakan merupakan kumpulan interaksi dari media sosial X terkait tiga pasangan calon presiden pada masa kampanye pemilihan presiden 2024 yang disediakan oleh Media Kernels Indonesia. Data disimpan dalam format .csv termasuk *username* yang telah dienkripsi, teks *tweet*, *time stamp*, jenis interaksi, dan beberapa informasi pendukung lainnya.

## 2.2 Data Cleaning

### 1. EDA (*Exploratory Data Analysis*)

Tahap ini berfokus pada analisa fitur-fitur atau kolom dari dataset yang diberikan oleh panitia, dengan tujuan untuk memahami karakteristik data. Metode analisis yang digunakan berupa analisis deskriptif sederhana seperti mencari nilai unik pada setiap kolom kategori, atau statistika sederhana pada kolom bertipe numerikal. Hal ini bertujuan untuk mempermudah pada proses pembersihan data seperti mengidentifikasi penyebab nilai unik yang terlalu banyak atau tak wajar. Selain itu, eksplorasi pada distribusi data juga dilakukan termasuk distribusi waktu terbentuk, distribusi lokasi, banyak teks unik, dan sebagainya.

### 2. Cleaning

Pada tahap ini, dilakukan penyelesaian permasalahan yang telah diidentifikasi dari eksplorasi data sebelumnya. Permasalahan umum pada data teks dari Twitter meliputi tidak konsistensi huruf kapital, stopword, singkatan dan kesalahan ketik, karakter khusus, nilai yang hilang, URL atau link, token-token khusus dari twitter seperti [RT] dan [RE], serta hashtag. Berikut langkah-langkah yang diambil dalam pembersihan data:

1. **Huruf Kapital:** Normalisasi huruf kapital dihindari sebab beberapa model transformer canggih seperti xlm-RoBERTa sudah dirancang untuk menangani sensitivitas huruf melalui tokenisasi berbasis sub-word (*WordPiece*) yang membantu menjaga konteks dan makna dari kata-kata tersebut seperti penandaan emosi, penandaan tempat, singkatan organisasi dan lain-lainnya.
2. **Stopword, singkatan dan kesalahan ketik:** Pada model kontekstual seperti transformer modern, umumnya telah dirancang dengan mekanisme atensi

untuk mengenali stopword. Penghapusan kata stopword dapat berisiko kehilangan konteks kalimat. Contohnya penghapusan kata seperti "tidak", "jangan", "dan", "di", dan sebagainya dapat menghilangkan informasi sentimen, struktur kalimat, pengenalan tempat, dan sebagainya. Untuk singkatan dan kesalahan ketik, walaupun jika diselesaikan dengan baik dapat meningkatkan performa model, Perlu diingat bahwa kombinasi dari volume data yang sangat besar dan kompleksitas waktu yang cukup tinggi dari metode konvensional seperti Levenshtein distance menyebabkan pendekatan ini tidak terskalakan.

3. **Karakter Khusus (*Special Characters*):** Karakter-karakter khusus yang akan disimpan hanya meliputi tanda baca, hashtag, dan token khusus seperti "<>", "[]". Karakter lainnya akan dihapus karena umumnya menimbulkan noise dan sering kali disebabkan oleh kesalahan format dari data aslinya, seperti kesalahan pengkodean pada emoji. Selain itu, penghapusan karakter khusus yang berulang, seperti "!!!" menjadi "!", juga dilakukan untuk menghemat jumlah token.
4. **Missing Value:** *Missing value* akan diisi placeholder seperti "0" atau "none" sesuai dengan jenis tipe datanya pada setiap kolom.
5. **URL, Special Token Tweet RE, RT dan mention:** Semua URL akan dihapus dari teks untuk mengurangi noise dan mengurangi jumlah token yang diproses agar beban komputasi tidak terlalu berat. Token RE (penanda reply) dan RT (penanda retweet) serta mention username akan diekstrak ke dalam fitur atau kolom baru untuk menyimpan informasi yang berpotensi penting untuk dianalisis seperti username yang direply.
6. **Hashtag:** Pada umumnya hashtag dalam teks tersebut mengandung informasi sentiment atau pun topik. Oleh karena itu, Hashtag dalam teks akan diekstrak ke dalam fitur atau kolom baru namun tetap dipertahankan dalam teks asli karena xlm-RoBERTa mampu menangani hashtag dengan baik.
7. **Inconsistency Label:** Label yang tidak konsisten pada suatu kolom kategoris seperti "Surabaya", "Sby", akan distandarisasi menggunakan library rapidfuzzy dan kamus yang disusun sendiri

### 2.3 Data labelling



*Data labelling* atau pelabelan data adalah proses untuk mengkategorikan *dataset* ke dalam kelompok kata tertentu yang memiliki makna dan informatif sehingga memberikan konteks yang akan dipelajari oleh model *machine learning* yang akan dibuat. Ada beberapa metode untuk melakukan *data labelling*, namun dalam penelitian ini menggunakan metode semi-otomatis. *Data labelling* dengan metode semi-otomatis merupakan pelabelan data yang menggunakan komputasi tingkat tinggi secara otomatis dalam penelitian ini menggunakan layanan *batch* API OpenAI yaitu GPT-4o dan pelabelan manual dalam penelitian ini menggunakan *software* Label-Studio. Metadata yang digunakan untuk pelabelan ditunjukkan pada file lampiran [1].

#### 2.4 Manual data validation

*Data validation* adalah memeriksa kualitas dan keakuratan data sebelum membuat model baru yang akan dilatih. *Data validation* secara manual dilakukan setelah melakukan *data labelling*, hal ini bertujuan untuk memastikan keakuratan dari pelabelan menggunakan API GPT-4o yang divalidasi dengan pelabelan secara manual pada saat *data labelling*. Pada alur metodologi yang ditunjukkan pada gambar 1 setelah melakukan *manual data validation* akan ada dua kondisi, yaitu jika hasilnya memuaskan maka akan dilanjutkan pada metode selanjutnya sesuai alur dan jika hasilnya tidak memuaskan maka akan dilakukan *data labelling* lagi sampai hasil dari *manual data validation* memuaskan.

#### 2.5 Data Preprocessing

*Data preprocessing* diperlukan untuk mengubah dataset menjadi format numerik yang dapat dipahami oleh model, sehingga dapat meningkatkan akurasi dan kinerja model yang dibuat. Tahapan *data preprocessing* yang dilakukan meliputi:

1. **Transformasi Kolom “*Candidate Mentions*”:** Kolom “*Candidate Mentions*” akan ditransformasi menjadi representasi One-Hot Encoding. Teknik One-Hot Encoding digunakan dalam klasifikasi multi-label karena memungkinkan model untuk mempelajari pola yang terkait dengan setiap kandidat secara independen. Setiap kandidat akan diwakili oleh vektor biner di mana nilai 1 menunjukkan keberadaan kandidat tersebut dan nilai 0 menunjukkan ketiadaan kandidat tersebut.
2. **Konversi Label Sentimen dan Topik:** Semua label pada kategori sentimen dan topik akan dikonversi menjadi bentuk numerik menggunakan label encoder.



3. **Tokenisasi Teks:** Teks akan dikonversi menjadi urutan numerik menggunakan Tokenizer yang disediakan oleh library XLMRobertaTokenizer. Proses tokenisasi ini sangat krusial karena memungkinkan model untuk memahami makna dan konteks dari data teks yang dimasukkan. Tokenisasi memecah teks menjadi token-token yang lebih kecil yang dapat dianalisis oleh model.

## 2.6 Fine-Tuning with XLM-RoBERTa

*Fine-tuning* adalah proses yang melibatkan model pra-terlatih yang digunakan untuk melatih lebih lanjut pada dataset tertentu. Dalam penelitian ini model pra-terlatih yang digunakan adalah XLM-RoBERTa. Sehingga model ini dipilih karena kemampuannya dalam memahami konteks dan makna teks dalam berbagai bahasa. Proses *fine-tuning* yang terstruktur ini memastikan bahwa model XLM-RoBERTa dapat bekerja secara optimal sesuai dengan data yang digunakan dalam penelitian ini. Adapun prosesnya seperti berikut:

1. **Data Splitting:** Dataset yang telah diproses sebelumnya dibagi menggunakan teknik Stratified K-Fold Cross-Validation untuk memastikan distribusi label yang seimbang di setiap lipatan.
2. **Model Training:** Melatih model menggunakan dataset yang lebih spesifik, pada tahapan ini peneliti memasukkan parameter pelatihan seperti jumlah epoch, ukuran batch, learning rate, dan parameter parameter lainnya yang diperlukan dalam penelitian.

## 2.7 Data Validation

Secara konsep data validation pada tahap ini sama dengan manual data validation pada metode sebelumnya, yaitu memeriksa kualitas dan keakuratan data yang dihasilkan oleh model yang telah dibuat. Dalam tahap ini data yang dihasilkan akan divalidasi dengan data pra-terlatih pada model XLM-RoBERTa, Setelah divalidasi, model akan dievaluasi untuk menampilkan persentase hasil prediksi model untuk memprediksi sentimen, kandidat, dan topik yang terkandung dalam *content* dataset. Evaluasi dilakukan menggunakan metrik seperti *accuracy*, *precision*, *recall*, *F1-score*, confusion matrix.

## III. HASIL DAN PEMBAHASAN

Pada bagian ini akan dipaparkan hasil *manual data validation* sebelum memaparkan hasil *fine-tuning* model pra-terlatih XLM-RoBERTa terhadap data sample dari dataset sebanyak 31.000 data yang performanya akan dievaluasi menggunakan

metrik *accuracy* dan *F1-score*. Setelah model dievaluasi dan hasilnya cukup memuaskan maka model inilah yang akan peneliti implementasikan pada dataset utama untuk memberikan label yang sesuai dengan kinerja dari model yang peneliti buat.

### 3.1 Hasil *manual data validation*

SKOR TOPICS															SKOR KANDIDAT					
	precision	recall	f1-score	support	SKOR Sentiment Jika Comparative dimerge											precision	recall	f1-score	support	
						precision	recall	f1-score	support											
0	0.87	0.87	0.87	606												0	0.79	0.69	0.74	178
1	0.91	0.95	0.93	117												1	0.95	0.94	0.95	355
2	0.98	0.76	0.86	84		1	0.84	0.83	0.84	431						2	0.93	0.94	0.93	561
3	0.94	0.94	0.94	48		2	0.87	0.96	0.91	961						3	0.98	0.98	0.98	618
4	0.93	0.85	0.89	65		3	0.85	0.36	0.51	113						4	0.74	0.95	0.83	21
5	0.91	0.89	0.90	56		4	0.77	0.68	0.73	295						5	0.82	0.96	0.89	28
6	0.78	0.89	0.83	457												6	0.61	1.00	0.76	22
7	0.84	0.83	0.84	206												7	0.93	0.82	0.87	17
8	0.89	0.65	0.75	161																
					micro avg	0.85	0.85	0.85	1800							micro avg	0.93	0.93	0.93	1800
					macro avg	0.84	0.71	0.75	1800							macro avg	0.85	0.91	0.87	1800
					weighted avg	0.85	0.85	0.84	1800							weighted avg	0.93	0.93	0.93	1800
					samples avg	0.85	0.85	0.85	1800							samples avg	0.93	0.93	0.93	1800
					Accuracy: 0.8494											Accuracy: 0.9289				
Accuracy: 0.8544																				

Gambar 2. Hasil *manual data validation*

Dengan pengambilan sampel sebanyak 1800 data lalu dilakukan *labelling* secara manual dan *labelling* menggunakan GPT-4o. Dari hasil validasi secara keseluruhan label/kelas diperoleh akurasi untuk klasifikasi topik sebesar 0.8544, klasifikasi sentimen sebesar 0.8494, dan klasifikasi kandidat sebesar 0.9289.

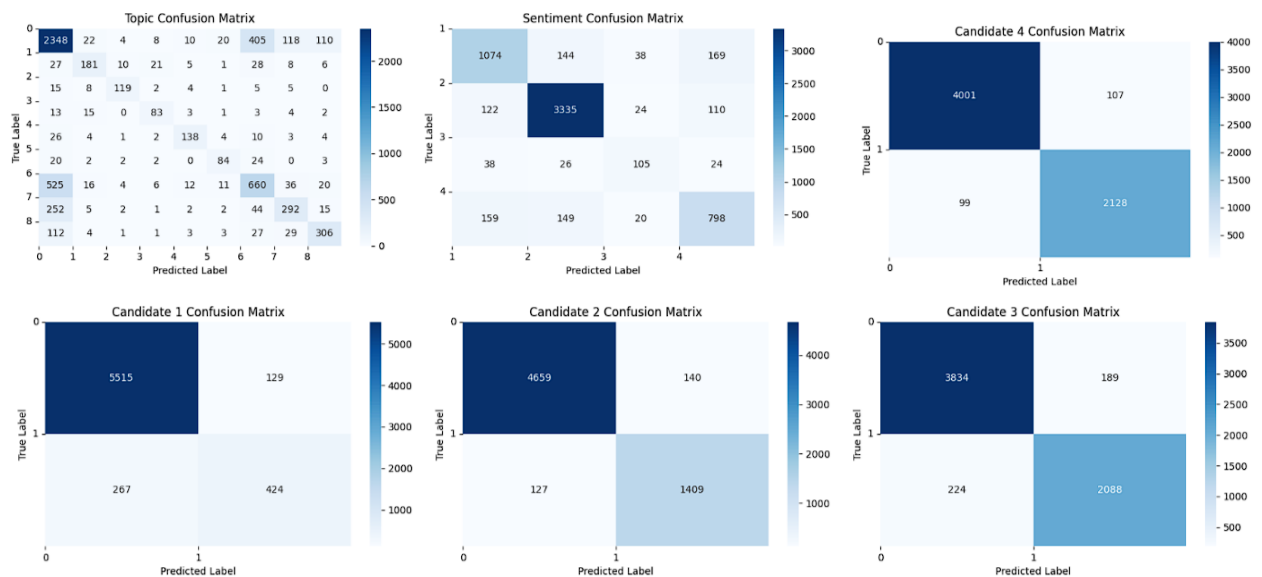
### 3.2 Hasil *fine-tuning* model

Epoch	Training Loss	Validation Loss	Candidate F1-Score	Sentiment Accuracy	Topic Accuracy
1	2.165700	2.112343	0.845313	0.805367	0.614680
2	1.781800	1.897998	0.861078	0.829992	0.654933
3	1.583300	1.870082	0.864221	0.838516	0.664720
4	1.410700	1.899978	0.864740	0.834886	0.663773

Tabel 1. Hasil *fine-tuning* model

Dari tabel 1 menunjukkan hasil *fine-tuning* model XLM-RoBERTa terhadap data sampel dengan 4 epoch/iterasi. Pada tabel 1 menunjukkan bahwa performa terbaik model untuk mengklasifikasikan sentimen terdapat pada epoch ke 3 dengan akurasi sebesar 0.838516, begitu juga dengan performa terbaik model untuk mengklasifikasikan topik dengan akurasi sebesar 0.664720, namun pada peforma terbaik model untuk mengklasifikasikan kandidat paslon dengan metrik F1-Score terdapat pada epoch ke 4 dengan score 0.86474. Untuk lebih detail persebaran hasil pelatihan model dapat dilihat

pada gambar 7 yang menunjukkan confusion matrix pada setiap label pada saat performa terbaik model.



Gambar 3. Confusion matrix performa terbaik model

Setelah melakukan *fine-tuning*, model yang memiliki performa terbaik secara keseluruhan yang akan tersimpan otomatis, pemilihan model terbaik ini juga melibatkan nilai *validation loss* dan *training loss*. Karena jika semakin kecil *validation loss* pada model maka model mencapai performa terbaik pada data validasi, yang merupakan indikator bahwa model tersebut mungkin dapat generalisasi dengan baik ke data baru. Di sisi lain, perlu diperhatikan pada *training loss* apabila *validation loss* mencapai titik terendah dan *training loss* menurun maka menunjukkan bahwa model dalam keadaan ideal yang artinya model tidak mengalami *overfitting* dan *underfitting*.

### 3.3 Model Implementation

Pada subbab ini model terbaik yang dipilih akan diimplementasikan ke dataset utama untuk melakukan klasifikasi yang sesuai dengan kinerja model pada saat pelatihan. Pada gambar 8 menampilkan hasil dari klasifikasi model setelah di *fine-tuning* terhadap 8 data pertama dalam dataset utama.

[18]:

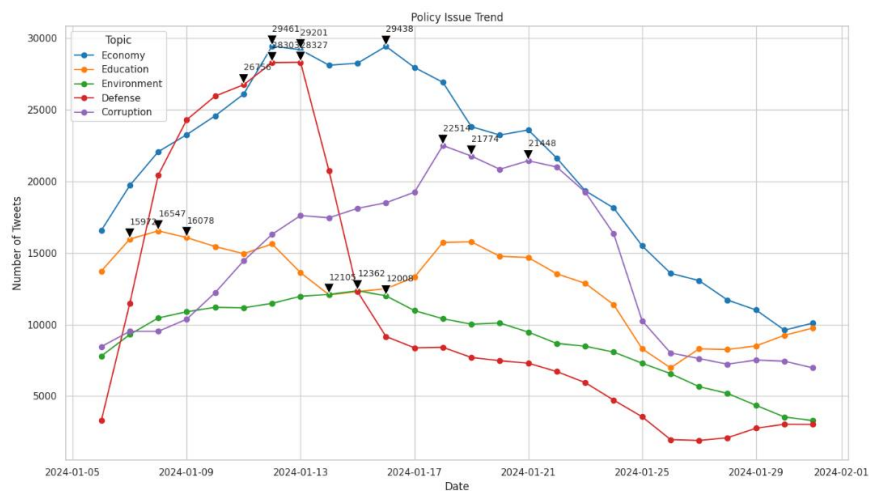
	index	content	predicted_candidate_mentions	predicted_sentiment	predicted_topic
0	0	Setelah debat ke tiga dan ke empat aku auto...	[3]	2	8
1	1	Ganjar Pranowo menghadiri Hajatan Rakyat Malan...	[3]	2	8
2	2	pak prabowo kalo kampanye selalu rame yaaah da...	[2]	2	8
3	3	Knp aktivis hny berani menantang ham sm pak p...	[2]	1	0
4	4	Bung buka mata dan telinga juga qulbi mu bi...	[0]	4	0
5	5	Prabowo Banks on Domestic Processing for Doubl...	[2]	2	1
6	6	Ganjar Pranowo Nah gitu den . kasih manfaat jg...	[1,3]	2	0
7	7	Wah seneng bgt Ganjar Gratiskan Int3rnet past...	[3]	2	1
8	8	Istirahat dulu pak nanti gemoynya ilang :	[0]	4	0

Gambar 4. Hasil klasifikasi model pada dataset

### 3.4 Interpretasi hasil

Berdasarkan hasil yang diperoleh dari klasifikasi pada tahap *model implementation*, berikut interpretasi dari hasil analisis dalam penelitian ini.

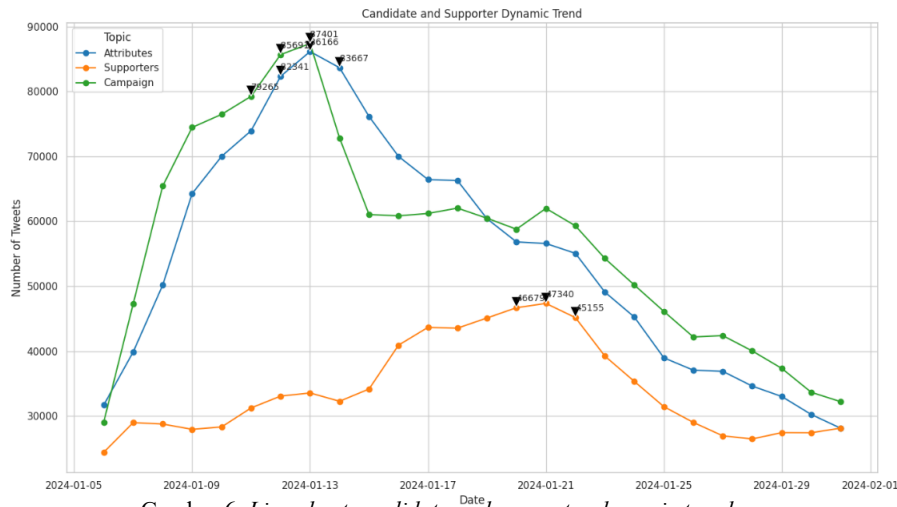
#### 3.3.1 Tren isu politik



Gambar 5. Line chart tren isu politik

Pada gambar 5 menunjukkan tren jumlah *tweet* per hari yang terkait dengan isu politik. Isu politik yang terdapat pada grafik, yaitu ekonomi, pendidikan, lingkungan, pertahanan, korupsi. Segitiga hitam pada gambar 9 menandakan 3 teratas jumlah *tweet* per hari yang berkaitan dengan masing masing topik.

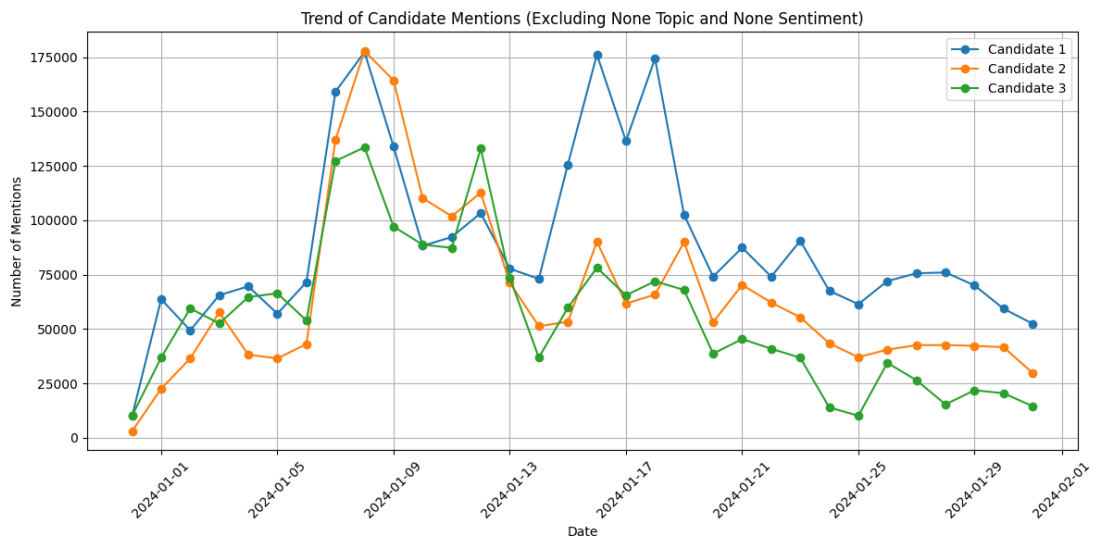
### 3.3.2 Candidate and supporter dynamic trend



Gambar 6. Line chart candidate and supporter dynamic trend

Pada gambar 6 menunjukkan tren jumlah *tweet* per hari yang terkait dengan isu *candidate and supporters*. Isu *candidate and supporter* yang terdapat pada grafik, yaitu *attribute*, pendukung, kampanye.

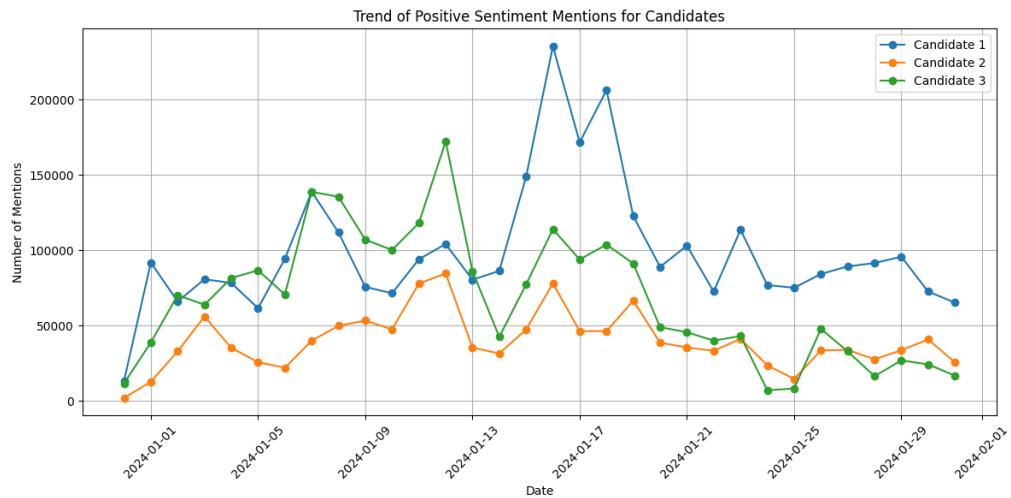
### 3.3.3 Candidate mention trend



Gambar 6. Line chart Candidate mention trend without none topic and none sentiment (neutral)

Pada gambar 6 menunjukkan tren jumlah *tweet* per hari yang mengandung topik dan sentimen terkait pasangan capres dan cawapres (*candidate mention*). Melalui gambar tersebut dapat disimpulkan bahwa popularitas atau perhatian terhadap pasangan urutan no 1 paling tinggi diantara kedua pasangan calon lainnya.

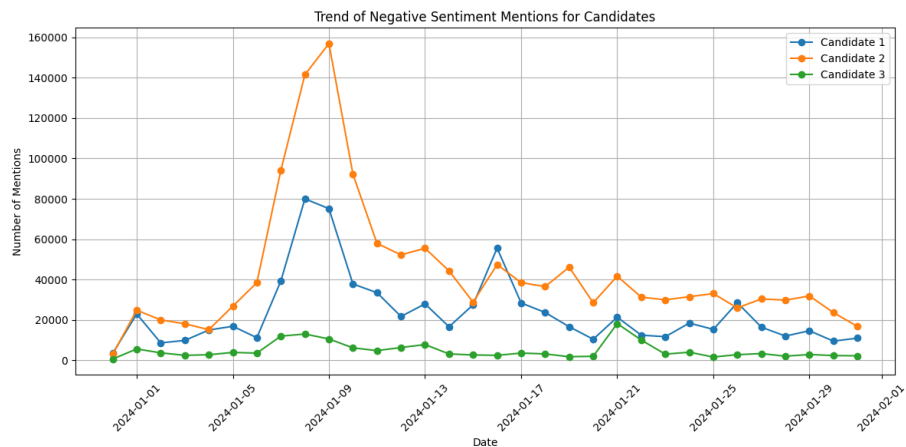
### 3.3.4 Sentimen positif setiap kandidat



Gambar 7. Line chart tren sentimen positif setiap kandidat

Pada gambar 7 menunjukkan tren jumlah *tweet* per hari yang mengandung sentiment positif terkait pasangan capres dan cawapres (*candidate mention*).

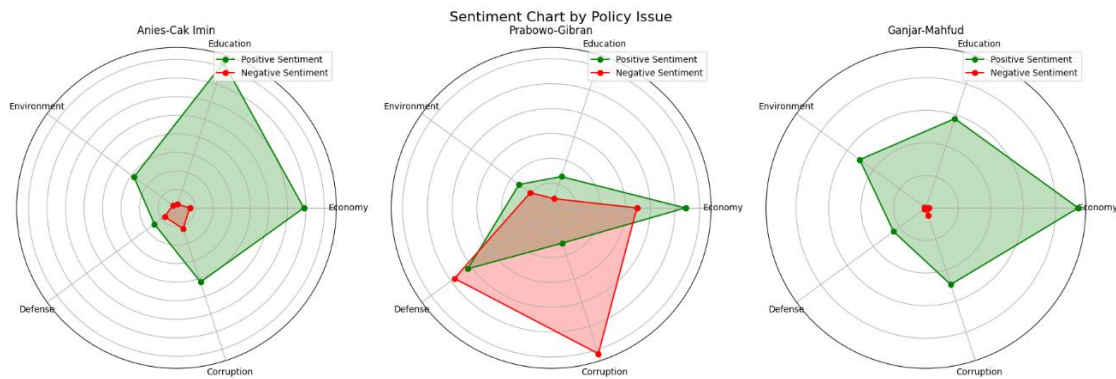
### 3.3.5 Sentimen negatif setiap kandidat



Gambar 8. Line chart tren sentimen negatif setiap kandidat

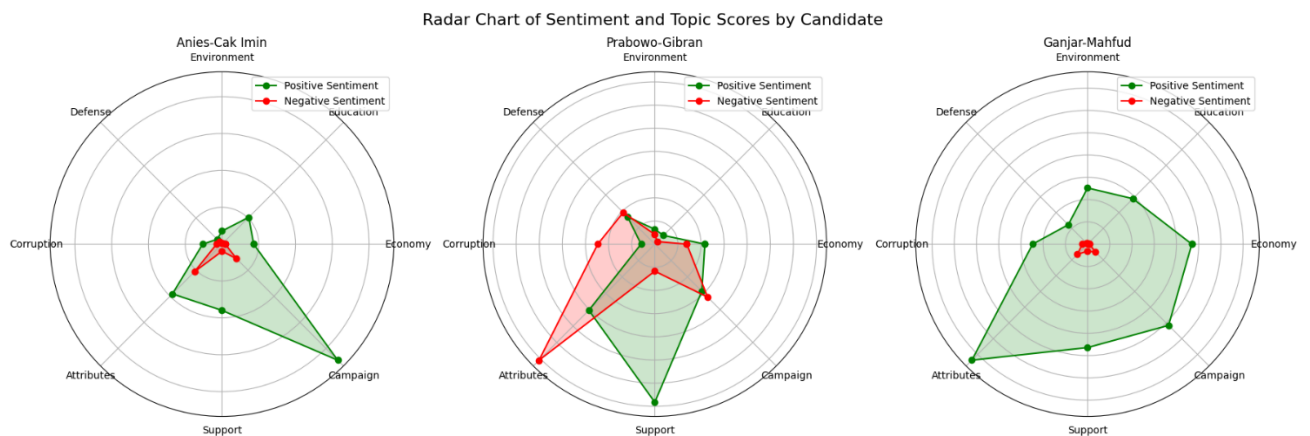
Pada gambar 8 menunjukkan tren jumlah *tweet* per hari yang mengandung sentiment negatif terkait pasangan capres dan cawapres (*candidate mention*). Melalui gambar tersebut peneliti **merekomendasikan** kepada kandidat 2 untuk meningkatkan citra melalui strategi komunikasi yang lebih transparan, seperti berkampanye dua arah antara audiens dan kandidat agar publik mengerti pola pikir dan pandangan kandidat 2 secara terbuka.

### 3.3.6 Analisis sentimen pada isu politik setiap kandidat



Gambar 9. Radar chart analisis sentimen pada isu politik

Pada gambar 9 menunjukkan analisis sentimen pada isu politik dari setiap kandidat. Kandidat 1 didominasi dengan sentimen positif yang berkaitan dengan pendidikan dan ekonomi. Kandidat 2 didominasi dengan sentimen negatif yang berkaitan dengan korupsi dan pertahanan. Kandidat 3 sangat didominasi oleh sentimen positif mengenai ekonomi dan pendidikan, kedua topik ini menjadi fokus utama yang dikaitkan kandidat 3. Maka peneliti **merekomendasikan** kandidat 2 untuk lebih mengutamakan isu korupsi dan pertahanan yang banyak dinilai negatif oleh publik dengan salah satunya membuat sebuah program yang berkaitan isu korupsi dan pertahanan.



Gambar 10. Radar chart analisis sentimen pada isu kandidat dan pendukung

Pada gambar 10 menunjukkan analisis sentimen pada isu kandidat dan pendukung dari setiap kandidat. Kandidat 1 didominasi dengan sentimen positif yang berkaitan dengan kampanye, topik ini yang menjadi perhatian khusus dari publik kepada kandidat 1 dikarenakan metode kampanye dari kandidat 1 dinilai memiliki terobosan baru dan berdampak positif banyak kalangan. Kandidat 2 didominasi dengan sentimen negatif dan positif, sentimen negatif berkaitan



dengan *attribute* dan sentimen positif berkaitan dengan pendukung, kedua topik ini memang menjadi perhatian dari publik karena *attribute* dari kandidat 2 yang dinilai publik kurang baik, namun pendukung dari kandidat inilah yang menaikkan citra dari kandidat 1. Kandidat 3 sangat didominasi oleh sentimen positif mengenai *attribute*, topik ini menjadi fokus utama yang dikaitkan oleh publik kepada kandidat 3. Maka peneliti **merekomendasikan** kepada kandidat 2 meningkatkan citra melalui strategi komunikasi yang lebih positif. Kandidat 2 sebaiknya fokus pada pendekatan yang dapat memperbaiki persepsi publik, seperti menunjukkan komitmen terhadap isu-isu yang menjadi *hot trending*.

#### IV. KESIMPULAN

##### 4.1 Kesimpulan

Data dibersihkan, termasuk normalisasi huruf kapital, stopword, singkatan, kesalahan ketik, karakter khusus, nilai hilang, *URL*, special token, dan hashtag. Sebuah pendekatan unik yang diajukan dalam penelitian ini adalah menggunakan GPT-4o melalui API untuk proses pelabelan. Pelabelan GPT-4o memiliki total akurasi diatas 90% pada indentifikasi kandidat dibahas, sentimen, hingga topik. Setelah divalidasi, pengambilan sampel data untuk *fine-tuning* menggunakan model berbasis XLM-RoBERTa sebagai model utama dalam proses klasifikasi melalui *prelabeled* GPT-4o.

Hasil *fine-tuning* menunjukkan bahwa model dengan performa terbaik secara keseluruhan memiliki akurasi sentiment sebesar 0.838516, akurasi topik sebesar 0.664720, dan F1-Score kandidat sebesar 0.864221. Analisis hasil klasifikasi model XLM-RoBERTa yang telah di *fine-tuning* menunjukkan bahwa topik yang paling banyak dibahas adalah seputar dinamika pendukung dan atribut kandidat, bukan kebijakan politik. Dengan ini, kami harapkan setiap kandidat memberikan prioritas utama dalam komunikasi dan peningkatan citra diri.

##### 4.2 Saran

Untuk penelitian kedepannya, peneliti menyarankan peningkatan sampel dan memperbanyak label topik. **Peningkatan jumlah sampel** merupakan saran dari peneliti untuk pengembangan lebih lanjut. Perlu diingat, peningkatan sampel juga mengakibatkan komputasi semakin tinggi dan mahal. Kemudian **memperbanyak label** topik bisa membantu model mengklasifikasikan topik dengan lebih rinci yang sedang dibahas dalam *content*.

## V. REFERENSI

KPU RI. (2022). *KEPUTUSAN KOMISI PEMILIHAN UMUM REPLUBLIK INDONESIA NOMO 21 TAHUN 2022*.

Maryanto, B. (2017). BIG DATA DAN PEMANFAATANNYA DALAM BERBAGAI SEKTOR. In *Media Informatika* (Vol. 16, Issue 2).

Astrianti Defretes, D., & Laga Kleden, K. (2023). EFEKTIVITAS PEMILIHAN UMUM SERENTAK TAHUN 2024. *Jurnal Hasil Penelitian*, 8, 49–50. <http://jurnal.untag-sby.ac.id/index.php/jhp17>

Wibowo, K. (2023). Kampanye Partisipatif Dalam Pemilihan Umum Tahun 2024. *Jurnal Pengawasan Pemilu*, 109–118.

Ismi Novia, R. (2020). STRATEGI KAMPANYE POLITIK PASANGAN JOKO WIDODO - MA'RUF AMIN DALAM PEMILIHAN PRESIDEN 2019 DI TWITTER. *Skripsi*.

Salman Farid, A. (2023). *PENGUNAAN MEDIA SOSIAL DALAM KAMPANYE POLITIK DAN DAMPAKNYA TERHADAP PARTISIPASI POLITIK DAN PERSEPSI PUBLIK* (Vol. 4, Issue 1).

Roihan, A., Abas Sunarya, P., & Rafika, A. S. (2019). IJCIT (Indonesian Journal on Computer and Information Technology) Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper. In *IJCIT (Indonesian Journal on Computer and Information Technology)* (Vol. 5, Issue 1).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). *Unsupervised Cross-lingual Representation Learning at Scale*. <http://arxiv.org/abs/1911.02116>

Basbeth, F., & Hatta Fudholi, D. (2024). *JURNAL MEDIA INFORMATIKA BUDIDARMA Klasifikasi Emosi Pada Data Text Bahasa Indonesia Menggunakan Algoritma BERT, RoBERTa, dan Distil-BERT*. 8(2), 1160–1170. <https://doi.org/10.30865/mib.v8i2.7472>

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. <http://arxiv.org/abs/1907.11692>

Jayadianti, H., Kaswidjanti, W., Utomo, A. T., Saifullah, S., Dwiyanto, F. A., & Drezewski, R. (2022). Sentiment analysis of Indonesian reviews using fine-tuning

IndoBERT and R-CNN. *ILKOM Jurnal Ilmiah*, 14(3), 348–354.  
<https://doi.org/10.33096/ilkom.v14i3.1505.348-354>