ANALISIS SENTIMEN KINERJA GUBERNUR LAMPUNG BERBASIS TEKS DAN *EMOJI* MENGGUNAKAN *INDOBERTWEET* DAN *XGBOOST*

(Skripsi)

Oleh

Moza Surya Putra

NPM 2217051014

Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Lampung
Bandar Lampung
2026

ANALISIS SENTIMEN KINERJA GUBERNUR LAMPUNG BERBASIS TEKS DAN *EMOJI* MENGGUNAKAN *INDOBERTWEET* DAN *XGBOOST*

Oleh

MOZA SURYA PUTRA

Skripsi Sebagai Salah Satu Syarat Untuk Mencapai Gelar
SARJANA KOMPUTER

Pada

Jurusan Ilmu Komputer
Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung

Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Lampung
Bandar Lampung
2026

ABSTRAK

ANALISIS SENTIMEN KINERJA GUBERNUR LAMPUNG BERBASIS TEKS DAN *EMOJI* MENGGUNAKAN *INDOBERTWEET* DAN *XGBOOST*

Oleh

MOZA SURYA PUTRA

Media sosial kini berperan sebagai sarana utama dalam penyampaian opini, kritik, dan saran masyarakat. Salah satu fenomena yang memerlukan analisis mendalam adalah persepsi masyarakat terhadap kinerja pemimpin daerah, khususnya Gubernur Lampung, yang tercermin melalui komentar pada media sosial. Komentar pada platform seperti *Instagram* dan *TikTok* bersifat beragam, tidak terstruktur, serta sering mengandung bahasa informal dan emoji, sehingga menyulitkan proses analisis sentimen. Penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap kinerja Gubernur Lampung berdasarkan komentar media sosial, mengukur kinerja model klasifikasi sentimen berbasis teks dan *emoji*, serta mengidentifikasi faktor linguistik yang memengaruhi terbentuknya sentimen. Metode yang digunakan adalah *feature-level fusion* antara teks dan emoji. Representasi teks diekstraksi menggunakan model *IndoBERTweet*, sedangkan representasi *emoji* dibentuk menggunakan model *Emojional*. Klasifikasi sentimen ke dalam tiga kelas, yaitu positif, negatif, dan netral, dilakukan menggunakan algoritma *XGBoost*. Data penelitian terdiri dari 34.127 komentar yang dikumpulkan melalui teknik *scraping* dari platform *Instagram* dan *TikTok*. Hasil penelitian menunjukkan bahwa penggabungan *IndoBERTweet* dan *Emojional* menghasilkan performa terbaik dengan nilai akurasi sebesar 89.247% dan *F1-score* sebesar 84.27%. **S**entimen positif mendominasi pada platform *Instagram* sebesar 60.87%, sedangkan sentimen negatif lebih dominan pada platform *TikTok* sebesar 53.09%. Analisis linguistik menunjukkan bahwa sentimen positif dipengaruhi oleh isu perbaikan infrastruktur jalan, program pemutihan pajak, dan pembebasan uang komite. Sebaliknya, sentimen negatif dipengaruhi oleh isu kerusakan jalan di beberapa wilayah, fluktuasi harga singkong, serta transparansi anggaran belanja pegawai.

Kata Kunci : Analisis Sentimen, IndoBERTweet, Emojional, XGBoost, Gubernur Lampung

ABSTRACT

SENTIMENT ANALYSIS OF THE GOVERNOR OF LAMPUNG'S PERFORMANCE BASED ON TEXT AND EMOJI USING INDOBERTWEET AND XGBOOST

By

MOZA SURYA PUTRA

Social media has become a primary medium for conveying public opinions, criticisms, and suggestions. One phenomenon that requires in-depth analysis is public perception of regional leaders' performance, particularly the Governor of Lampung, as reflected in social media comments. Comments on platforms such as Instagram and TikTok are diverse, unstructured, and often contain informal language and emojis, which complicates the sentiment analysis process. This study aims to analyze public sentiment toward the performance of the Governor of Lampung based on social media comments, evaluate the performance of text- and emoji-based sentiment classification models, and identify linguistic factors that influence sentiment formation. The proposed method employs feature-level fusion between textual and emoji features. Text representations are extracted using the IndoBERTweet model, while emoji representations are generated using the Emojional model. Sentiment classification into three classes positive, negative, and neutral is performed using the XGBoost algorithm. The dataset consists of 34,127 comments collected through scraping techniques from Instagram and TikTok. The results show that the integration of IndoBERTweet and Emojional achieves the best performance, with an accuracy of 89.247% and an F1-score of 84.27%. Positive sentiment dominates on the Instagram platform at 60.87%, while negative sentiment is more dominant on TikTok at 53.09%. Linguistic analysis reveals that positive sentiment is influenced by issues such as road infrastructure improvements, tax amnesty programs, and exemption of committee fees. In contrast, negative sentiment is influenced by road damage in several regions, fluctuations in cassava prices, and transparency issues related to personnel expenditure budgets.

Kata Kunci : Sentiment Analysis, IndoBERTweet, Emojional, XGBoost, Governor of Lampung

Judul Skripsi	: ANALISIS SENTIMEN KINERJA GUBERNUR LAMPUNG BERBASIS TEKS DAN *EMOJI* MENGGUNAKAN *INDOBERTWEET* DAN *XGBOOST*
Nama Mahasiswa	: Moza Surya Putra
Nomor Pokok Mahasiswa	: 2217051014
Program Studi	: Ilmu Komputer
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

MENYETUJUI

1. Komisi Pembimbing

Rahman Taufik, S.Pd, M. Kom
NIP. 19930627 202203 1 007

Ridho Sholehurrohman, M. Mat
NIP. 232111970128101

2. Ketua Jurusan Ilmu Komputer

Dwi Sakethi, S.Si., M.Kom.
NIP 196806111998021001

3. Ketua Program Studi Ilmu Komputer

Trisyanto, S.Kom, M.I.S., Ph.D.
NIP. 198104142005011001

MENGESAHKAN

1. Tim Penguji

Ketua : Rahman Taufik, S.Pd, M. Kom

Sekretaris : Ridho Sholehurrohman, M. Mat

Penguji Utama : Prof. Admi Syarif, Ph.D.

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP 197110012005011002

Tanggal Lulus Ujian Skripsi : 30 Januari 2026

PERNYATAAN

Saya yang bertanda tangan dibawah ini, menyatakan bahwa skripsi saya yang berjudul "**Analisis Sentimen Kinerja Gubernur Lampung Berbasis Teks dan *Emoji* Menggunakan *IndoBERTweet* Dan *XGBoost***" merupakan karya saya sendiri dan bukan karya orang lain. Semua tulisan yang tertuang di skripsi ini telah mengikuti kaidah penulisan karya tulis ilmiah Universitas Lampung. Apabila dikemudian hari terbukti skripsi saya merupakan hasil jiplakan atau dibuat orang lain, maka saya bersedia menerima sanksi berupa pencabutan gelar yang saya terima.

Bandar Lampung, 9 Februari 2026

Penulis,

Moza Surya Putra

NPM. 2217051014

SANWACANA

Puji syukur penulis ucapkan kehadirat Tuhan Yang Maha Esa, karena atas rahmat dan hidayah-Nya skripsi ini dapat diselesaikan.

Skripsi dengan judul "Analisis Sentimen Kinerja Gubernur Lampung Berbasis Teks dan *Emoji* Menggunakan *IndoBERTweet* Dan XGBoost " adalah salah satu syarat untuk memperoleh gelar sarjana di jurusan Ilmu Komputer di Universitas Lampung. Dalam Kesempatan ini penulis mengucapkan terimakasih kepada:

1. Ayah saya Surya Dharma dan ibu Emilia Mardani serta adik saya Dirga Surya Putra. selaku keluarga saya yang bekerja sangat keras untuk pendidikan saya dan segala kebutuhan saya;
2. Bapak Dr. Eng. Heri Satria, S.Si., M.Si. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam;
3. Bapak Dwi Sakethi, S.Si., M.Kom. selaku Ketua Jurusan Ilmu Komputer;
4. Ibu Yunda Heningtyas, M.Kom. selaku Sekretaris Jurusan Ilmu Komputer FMIPA Alam Universitas Lampung ;
5. Bapak Trisyanto S.Kom, M.I.S.,Ph.D. selaku Ketua Program Studi S1 Ilmu Komputer.
6. Bapak Prof. Admi Syarif, Ph.D. Selaku dosen Pembimbing Akademik dan Pembahas yang memberikan bimbingan, kritik, dan saran dalam proses pengerjaan skripsi ini;
7. Bapak Rahman Taufik, S.Pd., M.Kom. Selaku dosen Pembimbing I atas kesediannya memberikan bimbingan, kritik, dan saran dalam proses pengerjaan skripsi ini;
8. Bapak Ridho Sholehurrohman, M.Mat. Selaku dosen Pembimbing II atas kesediannya memberikan bimbingan, kritik, dan saran dalam proses pengerjaan skripsi ini;

9. Bapak dan Ibu Dosen Jurusan Ilmu Komputer FMIPA Universitas Lampung yang telah memberi ilmu dan membimbing saya selama kuliah;
10. Staff Jurusan Ilmu Komputer FMIPA Universitas Lampung.
11. Teman – teman tersayang yang selalu ada disisi saya selama masa perkuliahan.

Bandar Lampung, 9 Februari 2026

Penulis,

Moza Surya Putra
NPM. 2217051014

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

I. PENDAHULUAN

1.1. Latar Belakang

Media sosial dimanfaatkan sebagai saluran utama oleh masyarakat untuk mengekspresikan pandangan mereka. Media sosial diperkirakan digunakan oleh 130 juta orang secara aktif di indonesia atau sekitar 49% dari total penduduk di indonesia (Alfifa *et al*., 2025). Platform media sosial seperti *Instagram* dan *Tiktok* sangat populer di Indonesia, bahkan *Tiktok* dengan lebih dari 10 juta pengguna aktif telah menjadi salah satu platform berbagi video paling popular (Alfifa *et al*., 2025,Alfiah Zulqornain & Pandu Adikara, 2021). Berbagai opini publik tersaji melalui komentar di video atau postingan. Media sosial kini berperan sebagai saluran utama dalam menyampaikan opini, kritik, dan saran, menggantikan peran media tradisional yang cenderung kurang mudah dijangkau (Rizqilla *et al*., 2025).

Fenomena tersebut menimbulkan kebutuhan untuk memahami opini publik secara sistematis agar dapat diinterpretasikan secara objektif dan terukur (Rizqilla *et al*., 2025). Pendekatan yang dapat digunakan untuk memahami opini masyarakat salah satunya adalah analisis sentimen. Analisis sentimen merupakan pendekatan untuk mengenali arah pernyataan dalam suatu teks, apakah condong mendukung, menolak, atau berada di tengah, dengan mengacu pada makna yang dituturkan pengguna (Rivaldi *et al*., 2024). Permana *et al*. bahkan menegaskan bahwa analisis sentimen dapat digunakan untuk menguraikan komentar pengguna pada media sosial atas fenomena yang sedang ramai diperbincangkan (Permana *et al*., 2021). *Alexander et al*. menambahkan bahwa penambangan teks (*text mining*) memainkan peran penting dalam mengungkap pandangan masyarakat secara terstruktur melalui analisis sentimen (Alexander *et al*., 2023). Salah satu fenomena yang marak diperbincangkan dan memerlukan analisis mendalam adalah terkait kinerja pemimpin daerah, khususnya Gubernur Lampung.

Sebuah survei resmi mencatat bahwa pada 100 hari pertama kepemimpinan, tingkat kepuasan publik dalam layanan samsat mencapai sekitar 83,67%, yang mencerminkan adanya apresiasi terhadap sejumlah langkah kebijakan yang telah dijalankan. Namun di sisi lain, masih terdapat 34,20% responden yang merasa kurang puas dan 9,60% sangat tidak puas, terutama terkait dengan kondisi infrastruktur jalan (Lampung, 2025). Hal ini menunjukkan bahwa opini publik tidak bersifat tunggal, melainkan beragam dan bahkan saling bertolak belakang. Untuk memahami keragaman tersebut secara lebih mendalam, diperlukan pendekatan yang mampu membaca nuansa positif maupun negatif sekaligus, yaitu melalui analisis sentimen terhadap komentar masyarakat di media sosial.

Meskipun analisis sentimen di media sosial menawarkan banyak manfaat, prosesnya tidak selalu mudah. Unggahan di media sosial sering kali ditulis secara santai dan tidak mengikuti kaidah bahasa formal serta banyak yang menggunakan bahasa informal, singkatan, *emoji*. Hal ini membuat model analisis tradisional kesulitan memahami makna sebenarnya (Le *et al*., 2016).

Untuk menjawab tantangan diatas, pendekatan penelitian ini menggunakan *embedding* teks dengan *IndoBERTweet*. *IndoBERTweet* adalah model bahasa pra-latih berskala besar khusus teks *Twitter* bahasa Indonesia. Model ini memperluas *IndoBERT* dengan kosakata domain spesifik untuk *Twitter*, sehingga lebih efektif menangkap ciri khas teks media sosial (Koto *et al*., 2021). Penggunaan *IndoBERTweet* sebagai *embedding* terbukti meningkatkan akurasi model secara signifikan. *IndoBERTweet* meningkatkan akurasi model hingga 97,80% setelah dikombinasikan dengan *Word2Vec* (Azahra & Setiawan, 2023). Sementara itu, Indriani *et al* dalam penelitiannya menunjukkan bahwa *IndoBERTweet* memberikan performa yang lebih baik karena model tersebut dilatih menggunakan data media sosial berbahasa Indonesia yang sarat dengan bahasa informal, singkatan, kesalahan ejaan, serta ungkapan sehari-hari (Indriani *et al*., 2024) Hal tersebut menunjukkan *IndoBERTweet* merupakan pilihan yang tepat untuk digunakan sebagai metode *embedding teks* dalam analisis sentimen berbasis media sosial.

Selain teks, *emoji* juga menjadi sumber informasi emosional penting. Studi menunjukkan lebih dari 50% unggahan *Instagram* mengandung satu atau lebih *emoji* yang membantu memperjelas perasaan pengirim (Lou *et al*., 2024). Barry *et al* . mengembangkan *Emojional*, yaitu vektor *embedding* untuk *emoji* yang dilatih berdasarkan isi emosional setiap *emoji*. *Embedding* ini terbukti lebih akurat dalam analisis sentimen dibanding metode *embedding emoji* sebelumnya dalam beberapa aspek (Barry *et al*., 2022).

Al-Azani dan El-Alfy dalam penelitiannya melakukan eksperimen dengan menggabungkan representasi teks berbasis word *embedding* (*CBOW* dan *Skip-Gram*) dengan fitur *emoji* dalam bentuk sederhana, yaitu *emoji frequency*. Hasilnya menunjukkan bahwa kombinasi teks dan *emoji*, khususnya penggabungan *Text Skip-Gram* dengan *emoji frequency*, mampu meningkatkan akurasi klasifikasi sentimen secara signifikan dibandingkan jika hanya menggunakan teks atau *emoji* secara terpisah (Al-Azani & El-Alfy, 2021). Pada penelitian ini digunakan pendekatan serupa dengan peningkatan pada representasi teks, yaitu menggunakan *IndoBERTweet* sebagai *word embedding* yang lebih sesuai untuk bahasa Indonesia di media sosial, serta *Emojional* sebagai *emoji embedding* yang terbukti lebih akurat dalam analisis sentimen dibanding metode lain. (Koto *et al*., 2021). Dengan demikian, analisis sentimen diharapkan dapat menangkap nuansa bahasa dan penggunaan *emoji* secara lebih akurat.

Algoritma klasifikasi yang digunakan dalam penelitian ini adalah *Extreme Gradient Boosting* (XGBoost), yaitu algoritma berbasis pohon keputusan yang menerapkan pendekatan *ensemble learning* untuk memperbaiki kesalahan dari model sebelumnya. *XGBoost* dapat digunakan baik untuk regresi maupun klasifikasi (Yulistiani, 2024). *XGBoost* punya beberapa keunggulan, yaitu bisa mengurangi *overfitting* lewat regularisasi, berjalan lebih cepat karena mendukung pemrosesan paralel, mampu menangani data yang ada nilai hilangnya, dan sudah menyediakan *cross-validation.* Penelitian oleh Samih *et al*. menunjukkan bahwa penggunaan *embedding* bersama *XGBoost* mampu menghasilkan performa lebih baik dibanding metode tradisional seperti *TF-IDF*, sehingga algoritma ini efektif dalam mengolah representasi teks yang kompleks (Samih *et al*., 2023). Dengan kemampuan tersebut, *XGBoost* sesuai untuk dipadukan dengan I*ndoBERTweet* yang menghasilkan *embeddin*g khusus teks media

sosial bahasa Indonesia. Selain itu, penggabungan teks dengan fitur *emoji*, misalnya menggunakan pendekatan *emoji frequency* terbukti dapat meningkatkan akurasi analisis sentimen (Al-Azani & El-Alfy, 2021). Sinaga dan Agustian dalam penelitannya menunjukkan bahwa *XGBoost* mendapatkan hasil terbaik dibanding *decision tree* maupun semua metode yang digunakan dengan hasil akurasi 66% dan f1-score 57% (Sinaga & Agustian, 2022). Berdasarkan hal tersebut, penelitian ini akan menggunakan penggabungan teks dan *emoji* dengan *XGBoost* sebagai model klasifikasi.

Untuk menilai efektivitas dari sistem yang dibuat maka diperlukan metode evaluasi yang tepat seperti *precision* dan *recall* (Arora *et al*., 2016). *Recall* mengukur seberapa banyak sentimen yang relevan misalnya positif, negatif, atau netral berhasil ditemukan oleh model dari seluruh data yang seharusnya terdeteksi (Sadeli & Lawanda, 2023). Sementara itu, *precision* menunjukkan proporsi prediksi sentimen yang benar dibandingkan dengan seluruh prediksi yang dibuat oleh model untuk kategori tertentu (Martin & Nilawati, 2019). *Accuracy* adalah proporsi prediksi yang benar dari seluruh prediksi yang dilakukan (Sathyanarayanan, 2024). Untuk menentukan keseimbangan antara *precision* dan *recall* akan digunakan *F1 score* untuk menilanya(Cardova & Hermawan, 2023).

Untuk mengetahui faktor penyebab sentimen terhadap kinerja gubernur lampung maka akan digunakan teknik-teknik *Natural Languange Processing* (NLP). Teknik-teknik *NLP* yang akan digunakan adalah *n-gram*, *Part of Speech* (POS) *Tagging*, dan *Dependency parsing*. *N-gram* digunakan untuk mengidentifikasi pola kemunculan kata dan sekumpulan kata yang sering muncul dalam komentar pengguna, untuk mengetahui hubungannya dengan komentar (Surendro *et al*., 2024). *Part of Speech* (POS) *tagging* yang berfungsi untuk mengenali kategori gramatikal suatu kata dalam kalimat, seperti kata benda, kata kerja, atau kata sifat, dari suatu teks (Amien, 2023). *Dependency parsing* yaitu proses penting dalam *NLP* yang berfokus pada penentuan relasi gramatikal antar kata dalam suatu kalimat, sehingga membentuk representasi struktur sintaksis yang menunjukkan keterkaitan antar unsur kalimat tersebut(Kim & Ko, 2025).

Berdasarkan latar belakang yang telah dijelaskan, penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap kinerja Gubernur Lampung melalui

komentar pada platform *Instagram* dan *Tiktok* dengan memanfaatkan teks dan *emoji* sebagai sumber informasi. Model klasifikasi dibangun menggunakan *IndoBERTweet* untuk embedding teks, *Emoji*onal untuk embedding *emoji*, serta *XGBoost* sebagai algoritma klasifikasi guna mengukur performa model dalam mengklasifikasikan sentimen positif, negatif, dan netral. Selain mengevaluasi performa model dan menganalisis persebaran sentimen publik pada masing-masing platform dan *trend* tiap bulan, penelitian ini juga melakukan analisis linguistik terhadap komentar yang telah diklasifikasikan. Analisis linguistik dilakukan untuk mengidentifikasi penyebab sentiment terbentuk, sehingga penelitian ini tidak hanya menggambarkan kecenderungan opini publik, tetapi juga menjelaskan faktor linguistik yang memengaruhi sentimen masyarakat terhadap kinerja Gubernur Lampung.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana sentimen masyarakat terhadap kinerja Gubernur Lampung berdasarkan komentar di media sosial *Instagram* dan *Tiktok* ?
2. Bagaimana kinerja model *IndoBERTweet, Emojional*, dan *XGBoost* dalam melakukan klasifikasi sentimen terhadap komentar masyarakat di media sosial ?
3. Bagaimana hasil analisis linguistik pada sentimen terhadap kinerja Gubernur Lampung ?

1.3. Batasan Masalah

Untuk memfokuskan penelitian dan menghindari pembahasan yang terlalu luas, maka penelitian ini dibatasi pada:

1. Data yang dianalisis adalah komentar pada postingan yang terkait kinerja Gubernur Lampung di platform *Instagram* dan *Tiktok*.
2. Penelitian ini hanya menggunakan *XGBoost* sebagai model klasifikasi.
3. Penelitian menggunakan klasifikasi tiga kelas sentimen yaitu positif, negatif, dan netral.
4. Pengambilan komentar didasarkan pada postingan gambar dan video yang dipublikasikan dalam rentang waktu 21 Februari 2025 hingga 30 September 2025.
5. Analisis Linguistik hanya akan menggunakan teks alfabet yang ada di dalam komentar dan tidak termasuk simbol non-alfabet.

1.4. Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut :

1. Menganalisis sentimen publik antar platform *Instagram* dan *Tiktok* terhadap kinerja Gubernur Lampung berdasarkan hasil klasifikasi sentimen
2. Mengukur performa model klasifikasi sentimen yang menggunakan *IndoBERTweet* untuk teks, *Emojional* untuk *emoji*, serta *XGBoost* sebagai algoritma klasifikasi.
3. Mengetahui faktor penyebab sentimen terhadap kinerja Gubernur Lampung dengan analisis linguistik.

1.5. Manfaat Penelitian

Manfaat Penelitian ini adalah :

1. Memberikan pendekatan baru dalam penggunaan *embedding* teks dan *emoji* untuk analisis sentimen pada media sosial berbahasa Indonesia.
2. Menunjukkan efektivitas integrasi antara *word* dan *emoji embedding* yaitu *IndoBERTweet* dan *emojional* dengan algoritma *machine learning* klasik yaitu *XGBoos*t dalam klasifikasi sentimen komentar media sosial berbahasa Indonesia.
3. Menyediakan analisis sentimen publik antara platform *Instagram* dan *Tiktok* untuk mengidentifikasi persepsi masyarakat terhadap kinerja Gubernur Lampung.
4. Memberikan hasil analisis linguistik terkait faktor penyebab sentimen terhadap kinerja gubernur lampung, melalui teknik – teknik *Natural Languange Processing* (NLP), sehingga dapat menjadi referensi bagi peneliti lain atau pihak terkait dalam memahami opini publik.
5. Menjadi referensi dalam pengembangan alur kerja (*pipeline*) analisis sentimen berbasis teks dan *emoji* untuk media sosial berbahasa Indonesia.

II. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Tabel 1. Penelitian Terdahulu

No.	Penulis	Judul	Metode	Hasil
1.	(Al-Azani & El-Alfy, 2021)	*Early and Late Fusion of Emojis and Text to Enhance Opinion Mining*	Menggunakan dua strategi fusi yaitu *early fusion dan late fusion* untuk menggabungkan representasi teks dan *emoji* pada analisis sentimen.	- Fusi Tingkat Fitur (*Feature-Level Fusion* / *Early Fusion*) memberikan hasil akurasi tertinggi 83.74%. - Fusi Tingkat Skor (*Score-Level Fusion* / *Late Fusion*) memberikan hasil akurasi tertinggi 85.41%. - Fusi Tingkat Keputusan (*Decision-Level Fusion* / *Late Fusion*) memberikan hasil akurasi tertinggi

No.	Penulis	Judul	Metode	Hasil
				83.07%. Fusi Hibrida Multi-Level (*Multi-Level Hybrid Fusion*) memberikan hasil akurasi tertinggi 85.22%.
2.	(Liu *et al*., 2021)	*Improving sentiment analysis accuracy with emoji embedding*	Model *Chinese Emoji-embedding LSTM* berbasis (CEmo-LSTM) berbasis *Bi-LSTM* dikembangkan dan diuji menggunakan *dataset* yang terdiri dari 10.000 unggahan mikroblog berbahasa Tiongkok dari platform *Weibo.* Model ini melatih pengklasifikasi dengan menggabungkan fitur teks dan fitur *emoji*, di mana *emoji* diseleksi berdasarkan kategori *"Emoji Usage"*, sehingga hanya *emoji* yang berfungsi memperkuat ekspresi	- Model *CEmo-LSTM* mencapai akurasi tertinggi (sekitar 0.95) setelah diperkuat oleh filter *Emoji Usage*. Akurasi algoritma meningkat secara dramatis ketika dilatih hanya pada data yang sentimennya diperkuat oleh *emoji*.

No.	Penulis	Judul	Metode	Hasil
			sentimen dalam teks yang diproses oleh model.	
3.	(Nasrallah Taraf & T. Sadiq, 2024)	*Sentiment Analysis in Arabic Text and Emoji Using Deep Learning Methods*	Model yang digunakan dalam penelitian ini meliputi *AraBERT, DNN, CNN, LSTM, BERT*, serta kombinasi *CNN + LSTM* (dengan 2 dan 4 lapisan). Pengujian dilakukan dengan menggunakan *SpaCy word vector* dan *FastText* pada model *DNN, LSTM, CNN*, serta *CNN + LSTM* (2 dan 4 lapisan). Sementara itu, *AraBERT* tidak menggunakan *SpaCy word vector* atau *FastText*, melainkan menerapkan *emoji coding* sebagai	- *AraBERT* memberikan akurasi 84% menungguli model lainnya pada dataset *OMCD*. - *AraBERT* memberikan akurasi 95% mengungguli model lainnya pada *Arabic Sentiment Twitter Dataset*.

No.	Penulis	Judul	Metode	Hasil
			representasi fitur.	

Berbagai pendekatan telah digunakan untuk meningkatkan akurasi analisis sentimen berbasis teks dan *emoji*. Salah satu pendekatan dilakukan dengan strategi fusi, seperti *early fusion* dan *late fusion*, yang menggabungkan representasi teks dan *emoji* pada berbagai level. Hasilnya menunjukkan bahwa *Score-Level Fusion (Late Fusion)* dapat memberikan akurasi tertinggi hingga 85.41%, diikuti oleh metode *Multi-Level Hybrid Fusion* yang mencapai 85.22%, menandakan efektivitas penggabungan *multi-level* dalam memahami konteks emosi. Selain itu, penggunaan model berbasis *Bi-LSTM* yang mengombinasikan fitur teks dan *emoji* terbukti meningkatkan akurasi analisis hingga 0.95, terutama ketika hanya *emoji* yang relevan dengan ekspresi sentimen yang dipertahankan melalui proses seleksi. Di sisi lain, model *deep learning* seperti *AraBERT, DNN, CNN, LSTM, BERT*, dan kombinasi *CNN + LSTM* juga telah diuji dengan berbagai representasi vektor seperti *SpaCy word vector* dan *FastText*. Model yang memanfaatkan *emoji coding* secara langsung bahkan menunjukkan performa terbaik dengan akurasi mencapai 84% pada dataset umum dan 95% pada *dataset* media sosial berbahasa Arab.

Penelitian ini mengadopsi konsep-konsep dari pendekatan tersebut, khususnya gagasan *feature-level fusion* dalam penggabungan teks dan *emoji*, namun dikembangkan lebih lanjut dengan menggunakan *IndoBERTweet* sebagai representasi teks yang sesuai dengan bahasa informal Indonesia serta *Emojional* sebagai representasi *emoji* Hasil penggabungan keduanya kemudian diklasifikasikan menggunakan *XGBoost* untuk memperoleh performa yang stabil dan akurat, sehingga dapat memberikan hasil analisis sentimen yang lebih kontekstual terhadap data media sosial di Indonesia.

2.2. *Natural Languange Processsing* (NLP)

Natural Language Processing (NLP) merupakan cabang dari kecerdasan buatan (*AI*) yang berfokus pada bagaimana komputer dapat memahami dan menggunakan bahasa manusia. Tujuan utamanya adalah membuat komputer mampu berinteraksi dengan bahasa secara alami (Nurwanda *et al*., 2024). Beberapa teknik dasar dalam *NLP* antara lain *stemming*, yaitu proses mengubah kata ke bentuk dasarnya, misalnya kata "berlari" atau "berlarian" diubah menjadi "lari". Selain itu, terdapat *Part of Speech* (POS) *tagging* yang berfungsi untuk mengenali kategori gramatikal suatu kata dalam kalimat, seperti kata benda, kata kerja, atau kata sifat, sehingga komputer dapat lebih mudah memahami makna dari suatu teks (Amien, 2023), *n-gram yang* digunakan untuk mengidentifikasi pola kemunculan kata dan sekumpulan kata yang sering muncul dalam komentar pengguna, untuk mengetahui hubungannya dengan komentar (Surendro *et al*., 2024), dan *dependency parsing* yaitu proses penting dalam *NLP* yang berfokus pada penentuan relasi gramatikal antar kata dalam suatu kalimat, sehingga membentuk representasi struktur sintaksis yang menunjukkan keterkaitan antar unsur kalimat tersebut(Kim & Ko, 2025).

Berbagai cara dan teknik sudah dikembangkan untuk penelitian *NLP* bahasa Indonesia. *Machine Learning* membantu sistem mengenali jenis kata, serta digunakan untuk mengambil informasi dan menganalisis sentimen. Terjemahan mesin, termasuk *Neural Machine Translation* (NMT), membuat penerjemahan antarbahasa lebih tepat dengan mempelajari pola dari teks. Pendekatan berbasis konfiks digunakan untuk mengatasi masalah bentuk kata, misalnya dengan menghapus imbuhan agar kata kembali ke bentuk dasarnya (Amien, 2023).

Tantangan dalam *NLP* meliputi beberapa hal. Pertama, Bahasa Indonesia memiliki banyak imbuhan, prefiks, sufiks, dan afiks, sehingga kata-katanya sangat bervariasi dan membuat proses *stemming* atau *normalisasi* menjadi lebih kompleks. Kedua, ketersediaan dataset Bahasa Indonesia yang lengkap masih terbatas dibandingkan bahasa Inggris, sehingga menyulitkan penelitian *NLP* yang dapat direplikasi atau dibandingkan antar model. Selain itu, penggunaan kosakata informal, slang, dan

campuran bahasa di media sosial membuat komputer sulit memahami arti kata dengan tepat (Aji *et al*., 2022).

2.3. *Stanza*

Stanza merupakan sebuah *toolkit Natural Language Processing* (NLP) *open source* berbasis *Python* yang dikembangkan oleh tim dari *Stanford University*. Teknologi ini dirancang untuk menjembatani keterbatasan alat *NLP* sebelumnya yang seringkali hanya mendukung sedikit bahasa atau kurang akurat dalam memproses teks mentah. Keunggulan utama Stanza terletak pada kemampuannya memproses 66 bahasa manusia menggunakan sistem saraf (*neural pipeline*) yang utuh, mulai dari memecah kalimat hingga mengenali entitas penting seperti nama orang atau lokasi. Karena arsitekturnya bersifat *language-agnostic*, *Stanza* dapat mempelajari pola berbagai bahasa secara mandiri hanya dari data yang diberikan. *Stanza* mampu memproses teks mentah hingga menghasilkan anotasi linguistik yang komprehensif, mulai dari tokenisasi, lemmatisasi, *Part of Speech* (POS) *Tagging*, hingga *Named Entity Recongnition* (NER) dan analisis struktur sintaksis melalui *dependency parsing* (Qi *et al*., 2020).

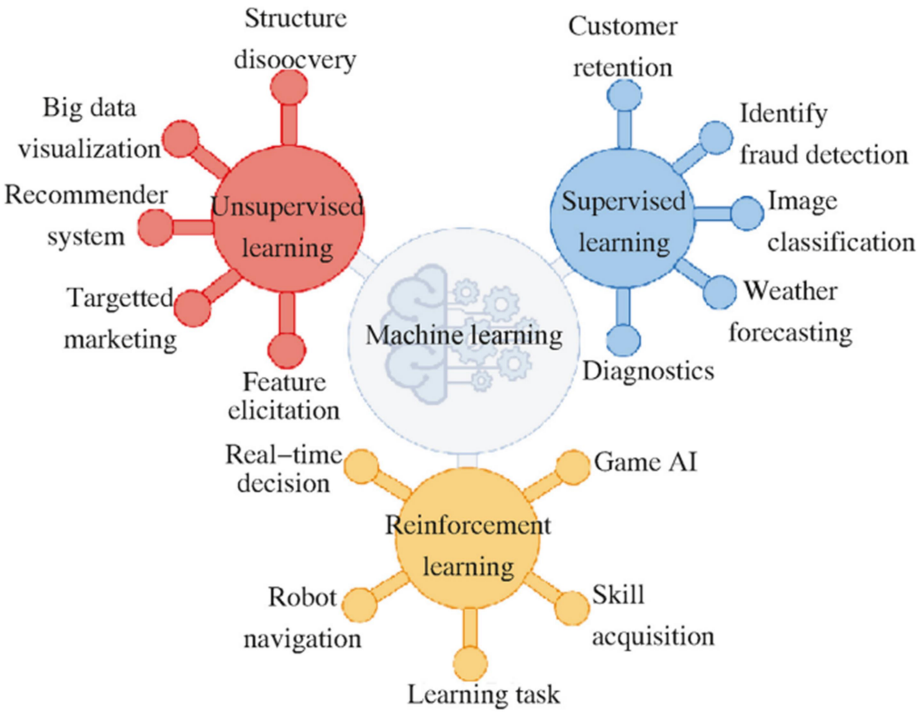
2.4. *Text Mining*

Text mining adalah metode untuk mengubah teks tidak terstruktur menjadi data yang lebih rapi dan mudah dianalisis. Prosesnya biasanya dimulai dari pengumpulan data teks, lalu dilanjutkan dengan pembersihan seperti menghapus tanda baca, URL, atau kata yang tidak penting, kemudian diubah ke dalam bentuk yang bisa dipahami oleh mesin melalui *feature extraction* seperti vektor kata. Dalam praktiknya, *text mining* sangat bergantung pada *Natural Language Processing* (NLP) karena banyak tahapannya, seperti tokenisasi, *stemming* atau *lemmatization*, penghapusan *stop words*, hingga memahami konteks, menggunakan teknik *NLP* agar teks bisa diolah secara otomatis oleh komputer (Astuti *et al*., 2024).

2.5. *Machine Learning*

Machine learning adalah cabang dari kecerdasan buatan yang memungkinkan komputer belajar dan berperilaku seperti manusia dengan meningkatkan kemampuannya melalui pengalaman. Sistem ini dirancang agar bisa mengambil keputusan sendiri dan menyesuaikan diri dengan perubahan tanpa perlu diprogram ulang. Proses *machine learning* membutuhkan data sebagai input untuk menemukan pola dalam kumpulan data besar, yang digunakan untuk melatih algoritma agar mampu melakukan analisis dengan tepat. Data dibagi menjadi data *training* untuk melatih model dan data *testing* untuk mengevaluasi kinerjanya saat menghadapi data baru (Retnoningsih & Pramudita, 2020).

Machine learning dibagi menjadi beberapa kategori berdasarkan cara model belajar dari data seperti yang bisa dilihat pada gambar 1. *Supervised learning* menggunakan data yang sudah berlabel sehingga model bisa mempelajari hubungan antara input dan output, cocok untuk tugas seperti klasifikasi, regresi, pengenalan wajah atau suara, tetapi membutuhkan banyak komputasi dan tidak bisa menemukan pola sendiri. *Unsupervised learning* bekerja dengan data tanpa label, di mana model mencoba menemukan pola atau kelompok berdasarkan fitur data, berguna untuk klasterisasi, analisis jaringan sosial, atau segmentasi pasar, meskipun hasilnya kadang sulit dievaluasi dan akurasinya lebih rendah. *Semi-supervised learning* memanfaatkan kombinasi data berlabel dan tak berlabel untuk mengurangi kebutuhan data beranotasi, namun akurasinya tidak selalu tinggi dan hasilnya bisa tidak stabil. *Reinforcement learning* belajar dari pengalaman melalui sistem hadiah dan hukuman dalam suatu lingkungan, digunakan pada prediksi lalu lintas, game, dan unggul dalam menyelesaikan masalah kompleks meski memerlukan banyak data dan komputasi (Pugliese *et al*., 2021).



Gambar 1. Klasifikasi Algoritma *Machine Learning*
(Pugliese *et al*., 2021)

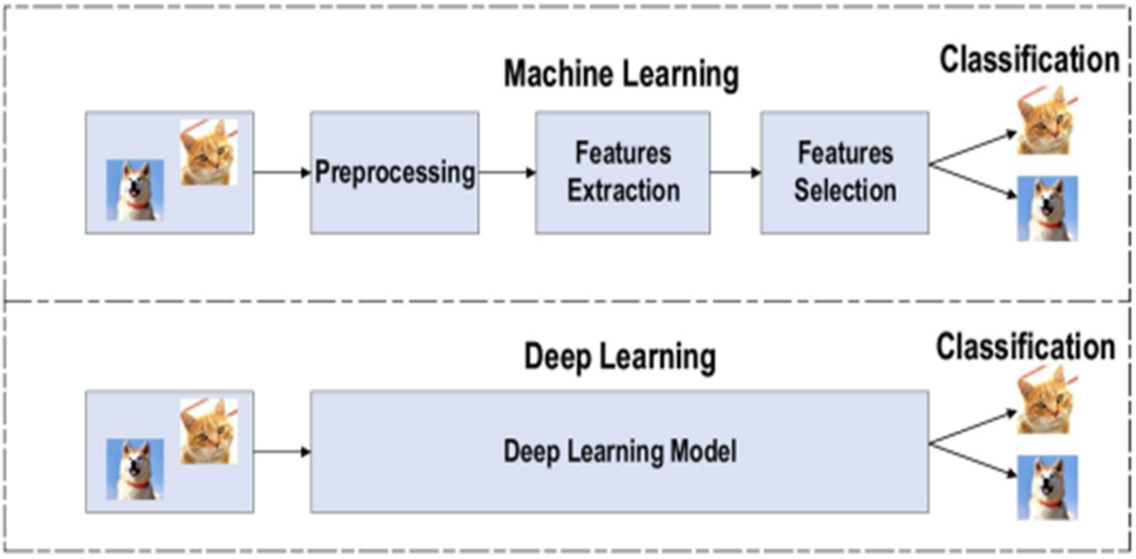
2.6. *Deep Learning*

Deep Learning (DL) adalah bagian dari *Machine Learning* (ML) yang mampu belajar dari data dalam jumlah besar tanpa perlu aturan manual, mirip dengan cara otak manusia memproses informasi. *DL* menggunakan jaringan saraf tiruan (ANN) dengan banyak lapisan, di mana setiap lapisan mengekstrak pola berbeda dari data. Terdapat beberapa jenis *DL*, seperti *CNN* untuk gambar, *RNN* dan *LSTM* untuk data urutan, serta *Transformer* untuk memahami konteks dalam teks. Berbeda dengan *ML* tradisional yang membutuhkan rekayasa fitur manual dan lebih cocok untuk data terstruktur, *DL* dapat menangani data mentah dan kompleks, sehingga sangat berguna dalam tugas seperti analisis sentimen dan pemrosesan bahasa alami (Alzubaidi *et al*., 2021).

Teknik *Deep Learning* umumnya dibagi menjadi tiga kategori utama yaitu *supervised* atau terawasi, *semi-supervised* atau sebagian terawasi, *unsupervised* atau tidak terawasi,

dan *deep reinforcement learning*. *Supervised learning* menggunakan data berlabel untuk melatih model, seperti *RNN*, *CNN*, dan *DNN*. *Unsupervised learning* bekerja tanpa label sehingga model belajar menemukan pola atau struktur tersembunyi dalam data. Di antara keduanya, *semi-supervised learning* memanfaatkan kombinasi data berlabel dan tidak berlabel, sedangkan *Deep Reinforcement Learning* belajar melalui interaksi dengan lingkungan untuk menemukan tindakan yang memberikan imbalan terbaik (Alzubaidi *et al*., 2021).

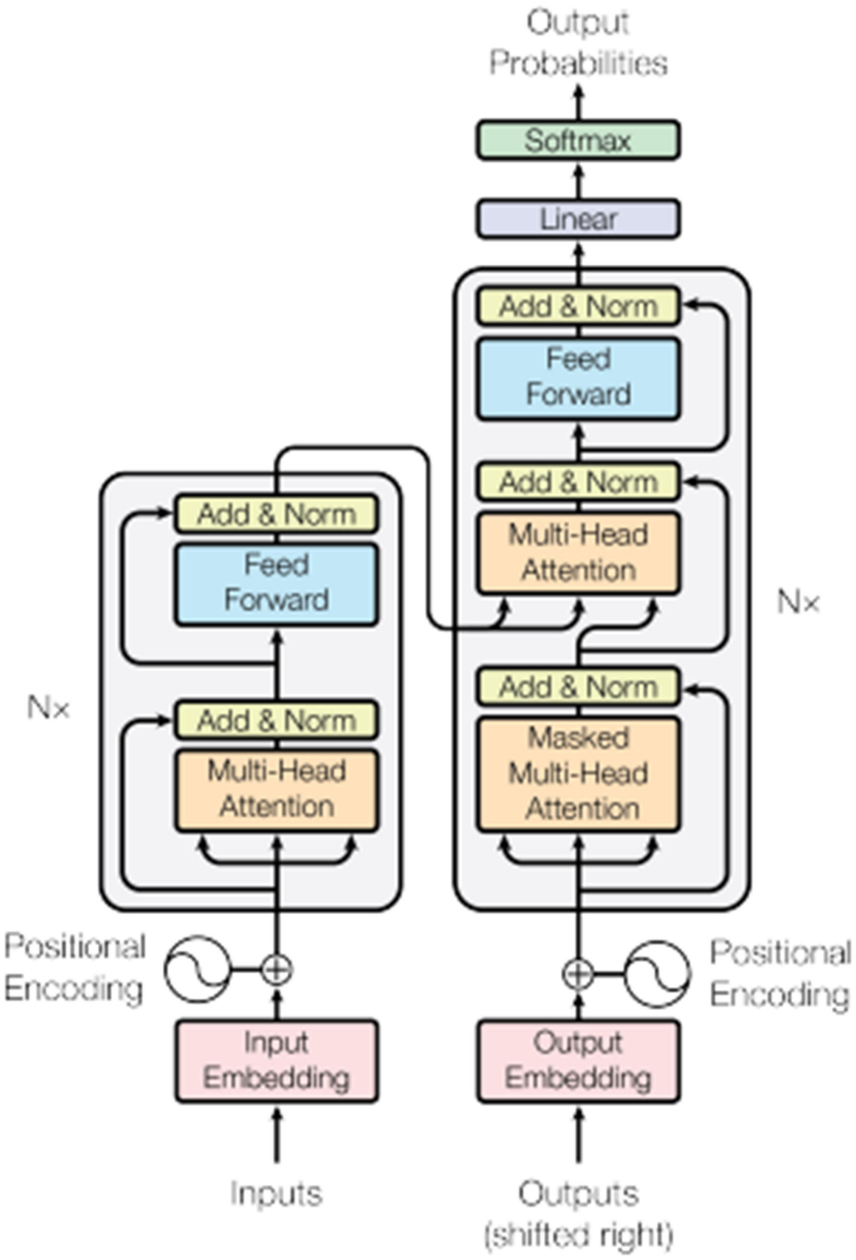
Perbedaan utama antara *Deep Learning* dan *Machine Learning* konvensional terletak pada cara fitur dipelajari seperti yang dapat dilihat pada gambar 2. Dalam *ML* tradisional, proses klasifikasi memerlukan langkah-langkah terpisah seperti pra-pemrosesan data, ekstraksi fitur, dan pemilihan fitur secara manual, yang sering membutuhkan banyak usaha manusia dan tergantung pada bidangnya. Sedangkan *Deep Learning* mengotomatisasi proses ini, sehingga jaringan dapat belajar dan mengklasifikasikan data secara langsung dalam satu langkah. Hal ini membuat *DL* lebih fleksibel dan lebih efektif dibandingkan ML tradisional, terutama untuk data dalam jumlah besar (Alzubaidi *et al*., 2021).



Gambar 2. Perbedaan *Machine Learning* dan *Deep Learning*
(Alzubaidi *et al*., 2021).

2.7. *Transformer*

Arsitektur *Transformer* merupakan salah satu inovasi penting dalam bidang pemrosesan bahasa alami karena hanya mengandalkan mekanisme *Attention* tanpa perlu menggunakan *Recurall Neural Netwok* (RNN) atau *Convolutional Neural Network* (CNN) yang kompleks. Dengan mekanisme ini, *Transformer* dapat memproses seluruh urutan data secara paralel sehingga pelatihan menjadi jauh lebih cepat dibandingkan model berbasis *RNN* yang bersifat sekuensial. Selain itu, *Transformer* juga mampu menangkap hubungan antar kata yang berjarak jauh dalam sebuah kalimat melalui *Attention*, sehingga lebih efektif dalam memahami konteks dibandingkan arsitektur sebelumnya (Vaswani *et al*., 2017). Keunggulan inilah yang kemudian dimanfaatkan dalam pengembangan berbagai model bahasa modern, salah satunya adalah *Bidirectional Encoder Representations from Transformers* (BERT) yang menggunakan arsitektur *Transformer* untuk menghasilkan representasi teks yang lebih kontekstual (Devlin *et al*., 2019). Pembahasan mengenai arsitektur *Transforme*r pada subbab 2.8.1 hingga 2.8.5 merujuk pada penelitian Vaswani *et al*. yang diperkenalkan dalam penelitiannya (Vaswani *et al*., 2017) .



Gambar 3. Arsitektur Model *Transformer*

(Vaswani *et al*., 2017)

2.7.1. *Encoder* dan *Decoder*

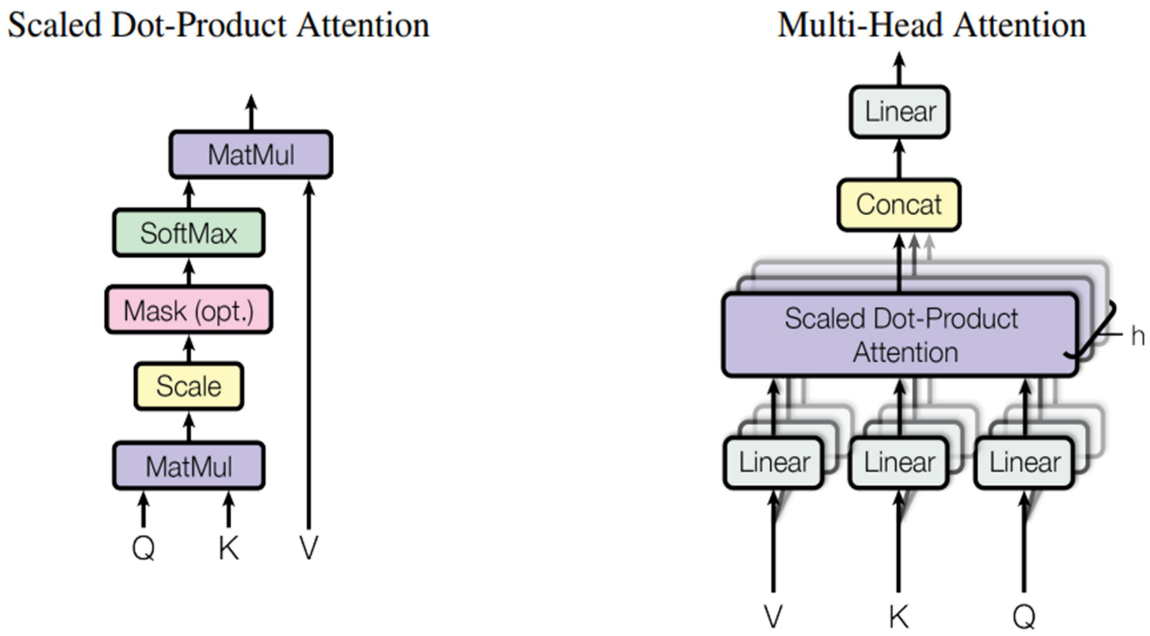
Pada bagian *encoder*, arsitektur *Transformer* terdiri dari 6 lapisan yang sama. Setiap lapisan memiliki dua sub-lapisan, yaitu mekanisme *multi-head self-attention* dan *feed-forward network*. Setiap sub-lapisan dibungkus dengan *residual connection* dan dilanjutkan dengan *layer normalization*, sehingga *output* setiap lapisan dinotasikan sebagai :

$$LayerNorm(x + sublayer(x)) \quad (1)$$

Agar konsisten, semua sub-lapisan dan *layer embedding* menghasilkan output dengan dimensi 512. Sedangkan pada bagian *decoder*, strukturnya juga terdiri dari 6 lapisan yang sama, tetapi memiliki tambahan satu sub-lapisan. Jadi selain dua sub-lapisan seperti *encoder*, terdapat sub-lapisan ketiga berupa *multi-head attention* yang mengambil informasi dari hasil *encoder*. Sama seperti *encoder*, setiap sub-lapisan di *decoder* juga menggunakan *residual connection* dan *layer normalization*. Bedanya, pada *self-attention* di *decoder* diterapkan masking untuk mencegah posisi tertentu melihat kata setelahnya. Dengan cara ini, prediksi pada posisi *i* hanya bisa menggunakan informasi dari posisi sebelumnya, sehingga model bisa melakukan prediksi secara berurutan.

2.7.2. *Attention*

Fungsi *attention* dapat dijelaskan sebagai proses pemetaan antara sebuah *query* dengan sekumpulan *key*-*value pair* untuk menghasilkan sebuah output. Baik *query*, *key*, *value*, maupun *output* semuanya direpresentasikan dalam bentuk vektor. *Output* tersebut dihitung sebagai penjumlahan berbobot dari nilai (*values*), di mana bobot tiap nilai ditentukan oleh tingkat kesesuaian (*compatibility*) antara *query* dan *key* yang bersangkutan.



Gambar 4. Ilustrasi (a) *Scaled Dot-Product Attention* (b) *Multi-Head Attention* (Vaswani *et al*., 2017)

Jenis *attention* yang dipakai adalah *Scaled Dot-Product Attention*, yang ditunjukkan pada Gambar 4 bagian (a). Masukan yang diproses terdiri dari *query* dan *key* dengan dimensi $\boldsymbol{d_k}$, serta *value* dengan dimensi $\boldsymbol{d_v}$. Prosesnya dilakukan dengan menghitung hasil perkalian titik (*dot product*) antara *query* dengan semua *key*, lalu hasilnya dibagi dengan$\sqrt{\boldsymbol{d_k}}$ dan diterapkan fungsi *softmax* untuk mendapatkan bobot dari masing-masing *value*. Dalam penerapannya, perhitungan ini dilakukan secara bersamaan untuk sekumpulan *query* yang direpresentasikan dalam matriks *Q*, sedangkan *key* dan *value* direpresentasikan dalam matriks *K* dan *V*. Hasil akhirnya berupa matriks output yang dirumuskan sebagai berikut:

$$\boldsymbol{Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V} \qquad (2)$$

Multi- Head Attention yang ditunjukkan pada gambar 4 (b) memiliki perbedaan dengan *Scaled Dot-Product Attention*. Daripada hanya menggunakan satu fungsi *attention* dengan *query*, *key*, dan *value* berdimensi d_{model}, pendekatan yang digunakan adalah memproyeksikan *query*, *key*, dan *value* masing-masing sebanyak h kali ke dalam dimensi yang lebih kecil, yaitu d_k, d_k, dan d_v, melalui proyeksi linear yang dipelajari

secara terpisah. Pada setiap versi proyeksi tersebut, fungsi *attention* dijalankan secara paralel sehingga menghasilkan *output* berdimensi d_v. Hasil dari semua head ini kemudian digabungkan (*concatenate*) dan diproyeksikan kembali untuk menghasilkan nilai akhir.

Secara analogi, *multi-head attention* dapat dibayangkan seperti memiliki beberapa orang pembaca (*head*) yang fokus pada aspek berbeda dari teks yang sama. Satu orang mungkin memperhatikan hubungan antar kata di awal kalimat, sementara orang lain fokus pada akhir kalimat. Ketika semua hasil pengamatan mereka digabungkan, model memperoleh pemahaman yang lebih kaya dibandingkan jika hanya ada satu pembaca. Secara umum Multi-Head Attention dirumuskan sebagai berikut :

$$Multihead(Q,K,V) = Concat(head_1,\dots,head_h)W^O where\ \ head_i \qquad (3)$$
$$= Attention\ (QW_i^Q, KW_i^k, VW_i^V)$$

2.7.3. *Position-wise Feed-Forward Networks*

Setiap lapisan pada *encoder* dan *decoder* selain memiliki mekanisme *attention* juga dilengkapi dengan *feed-forward network* (FFN). Jaringan ini bekerja dengan cara yang sama pada setiap posisi token, yaitu melalui dua transformasi linear yang dipisahkan oleh aktivasi *ReLU*. Secara sederhana, proses ini berfungsi untuk mengolah informasi dari hasil *attention* secara lebih mendalam dengan memperluas dimensi internal menjadi lebih besar (d_{ff} = 2048) kemudian diproyeksikan kembali ke dimensi semula (d_{model}= 512). Dengan kata lain, FFN bertugas memperkaya representasi informasi sebelum diteruskan ke lapisan berikutnya. Rumusnya sebagai berikut :

$$FFN\ (x) = \max(0, xW_1 + b_1)\, W_2 + b_2 \qquad (4)$$

2.7.4. *Embedding* dan *Softmax*

Model ini mengubah kata-kata menjadi *embedding*, yaitu representasi vektor agar komputer bisa memahami maknanya. Vektor ini kemudian diproses oleh *decoder* untuk menebak kata berikutnya. Hasil *decoder* masuk ke *layer linear* dan *softmax*, di mana *softmax* berfungsi mengubah hasil prediksi menjadi probabilitas setiap kata, supaya model tahu kata mana yang paling mungkin muncul. Untuk lebih efisien, model memakai bobot yang sama antara tempat membuat vektor (*embedding*) dan tempat menebak kata (*linear layer*), dan nilai vektor dikalibrasi supaya tidak terlalu besar atau kecil dengan dikalikan dengan $\sqrt{d_{model}}$ supaya hasil representasinya tetap seimbang.

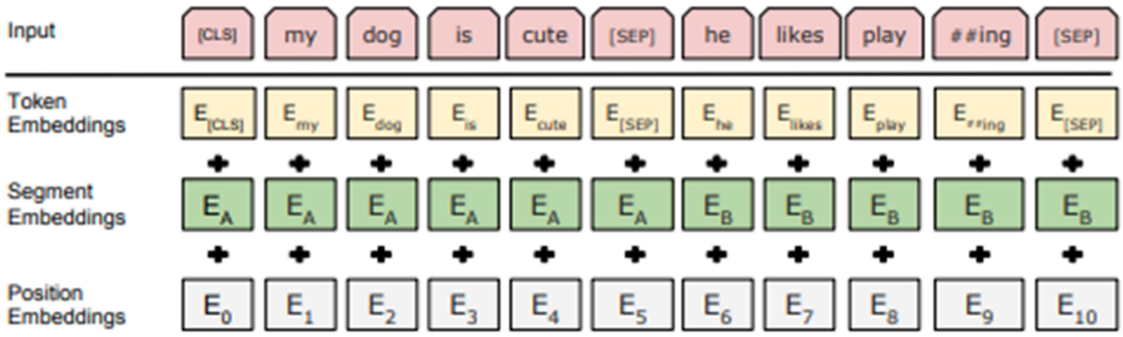
2.7.5. *Positional Encoding*

Positional encoding adalah teknik yang digunakan pada *Transformer* untuk memberi informasi urutan posisi token dalam sebuah kalimat, karena model ini tidak menggunakan *RNN* maupun *CNN* yang secara alami mempertahankan urutan. Cara kerjanya adalah dengan menambahkan vektor posisi ke embedding token, di mana setiap dimensi dihitung menggunakan fungsi sinus dan cosinus dengan frekuensi berbeda. Dengan demikian, setiap token tidak hanya memiliki representasi makna, tetapi juga informasi tentang posisinya dalam urutan, sehingga model dapat memahami hubungan relatif antar kata dalam kalimat.

2.8. *Bidirectional Encoder Representations from Transformers* (BERT)

Bidirectional Encoder Representations from Transformers (BERT) adalah model bahasa yang dikembangkan oleh *Google* untuk menghasilkan pemahaman teks secara dua arah (bidireksional). Artinya, *BERT* mampu memahami konteks sebuah kata dengan melihat kata di sebelah kiri maupun kanan secara bersamaan, bukan hanya salah satunya. Secara struktur, *BERT* dibangun menggunakan *multi-layer bidirectional*

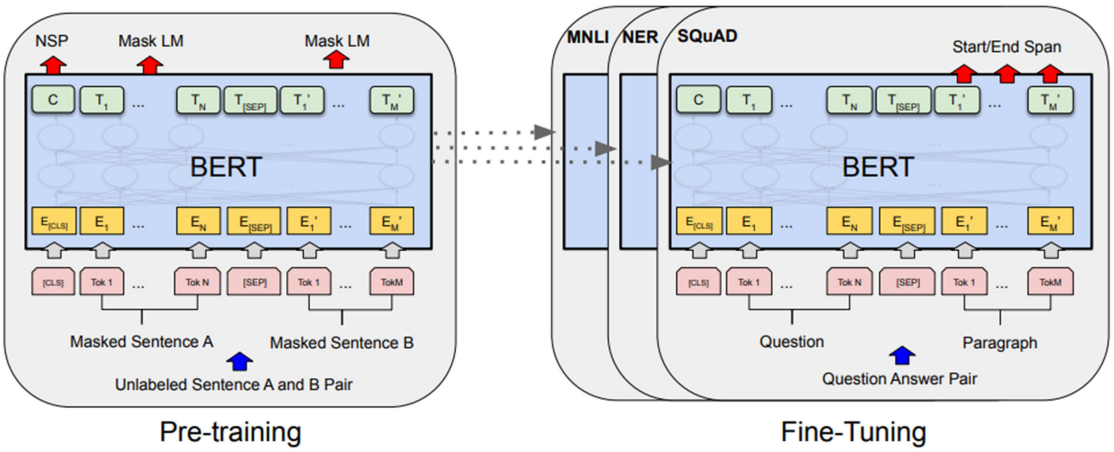
Transformer encoder yang diadaptasi dari arsitektur asli *Transformer Encoder*. Representasi input *BERT* ditunjukkan pada Gambar 5 dapat berupa satu kalimat maupun pasangan kalimat, misalnya *Question* dan *Answer*. Input ini dibentuk dengan menjumlahkan tiga komponen utama, yaitu *Token Embeddings* yang dihasilkan menggunakan *WordPiece* dengan kosakata sekitar 30.000 *token*, *Segment Embeddings* yang membedakan dua kalimat atau segmen, serta *Position Embeddings* yang menunjukkan urutan posisi kata dalam kalimat. Selain itu, *BERT* menggunakan token khusus [CLS] di awal urutan untuk keperluan klasifikasi, serta [SEP] sebagai pemisah antar kalimat (Devlin *et al*., 2019).



Gambar 5. Representasi Input *BERT*
(Devlin *et al*., 2019).

Tahap *pre-training* ditunjukkan pada Gambar 6 bagian kiri dilakukan dengan melatih *BERT* menggunakan dua sumber data besar, yaitu *BooksCorpus* yang berisi sekitar 800 juta kata dan *Wikipedia* bahasa Inggris dengan 2,5 miliar kata. Pada tahap ini, *BERT* belajar memahami bahasa melalui dua tugas utama. Pertama, *Masked Language Model* (MLM), yaitu dengan menyembunyikan sekitar 15% kata dalam kalimat secara acak, lalu meminta model menebak kembali kata yang hilang dengan melihat konteks di kiri dan kanan. Dalam proses ini, 80% kata yang disembunyikan diganti dengan *token* [MASK], 10% diganti dengan kata acak, dan 10% sisanya tetap dibiarkan. Kedua*, Next Sentence Prediction* (NSP), yaitu tugas untuk menebak apakah suatu kalimat benar-benar merupakan lanjutan dari kalimat sebelumnya (*IsNext*) atau hanya kalimat acak

(*NotNext*). Kedua jenis latihan ini membuat *BERT* mampu memahami arti kata sekaligus hubungan antar kalimat dengan lebih baik (Devlin *et al*., 2019).



Gambar 6. Prosedur *Pre-Training* dan *Fine-Tuned* pada *BERT* (Devlin *et al*., 2019).

Setelah tahap *pre-training* selesai, model *BERT* dapat digunakan untuk tahap *fine-tuning* ditunjukkan pada gambar 6 bagian kanan. Pada tahap ini, hasil pelatihan awal dari *pre-training* dijadikan sebagai dasar, lalu disesuaikan kembali menggunakan data berlabel sesuai dengan tugas yang akan dikerjakan. Proses *fine-tuning* ini cukup sederhana karena *BERT* hanya memerlukan tambahan satu lapisan *output* agar bisa digunakan pada berbagai macam tugas (Devlin *et al*., 2019).

BERT tersedia dalam dua ukuran utama. *BERT-Base* terdiri dari 12 lapisan *encoder* dengan ukuran vektor 768, 12 *attention heads*, dan total sekitar 110 juta parameter. Sementara itu, *BERT-Large* lebih besar dengan 24 lapisan *encoder*, ukuran vektor 1024, 16 *attention heads*, dan total sekitar 340 juta parameter. Saat pertama kali diperkenalkan, *BERT* berhasil mencapai performa terbaik pada 11 tugas *NLP*. Beberapa di antaranya adalah skor *GLUE* sebesar 80,5% dan nilai *F1* pada *SQuAD v1.1* sebesar 93,2. Hasil ini membuktikan bahwa *BERT* menjadi salah satu pencapaian penting dalam perkembangan model bahasa modern, karena dapat digunakan secara luas pada berbagai jenis tugas hanya dengan penyesuaian yang minimal (Devlin *et al*., 2019).

2.9. *IndoBERTweet*

I*ndoBERTweet* merupakan model bahasa pra-latih pertama yang khusus dibuat untuk data *Twitter* berbahasa Indonesia. Model ini dikembangkan dengan memanfaatkan *IndoBERT* sebagai dasar, lalu menambahkan kosakata baru yang sering muncul di *Twitter* namun tidak ada dalam *IndoBERT*. Masalah utama yang dihadapi adalah banyak kata khas media sosial yang tidak dikenali dengan baik oleh model umum. Untuk mengatasi hal ini, peneliti mencoba beberapa cara menginisialisasi kosakata baru, dan hasil terbaik diperoleh dengan merata-ratakan *embedding subword* dari *IndoBERT*. Cara ini terbukti lebih efisien dibanding metode acak maupun proyeksi dari *word2vec* atau *fastText*, karena membuat proses pelatihan lebih cepat dan hasilnya lebih baik (Koto *et al*., 2021).

Data yang digunakan untuk melatih *IndoBERTweet* diambil dari 26 juta *tweet* berbahasa Indonesia yang dikumpulkan selama satu tahun penuh, dengan total sekitar 409 juta kata. *Dataset* tersebut mencakup topik-topik besar seperti ekonomi, kesehatan, pendidikan, dan pemerintahan. Dari data ini dibuat kosakata baru sebanyak hampir 32 ribu *token* menggunakan metode *WordPiece*. Arsitektur model *IndoBERTweet* masih mengikuti *IndoBERT*, yaitu *transformer encoder* dengan 12 lapisan, *12 attention head*, dan ukuran *hidden* 768, tetapi disesuaikan untuk teks *Twitter* dengan panjang maksimal 128 token. Pelatihannya menggunakan metode *masked language modeling*, baik dengan membangun model dari nol maupun dengan melakukan penyesuaian dari *IndoBERT* yang sudah ada (Koto *et al*., 2021).

Kinerja *IndoBERTweet* diuji menggunakan tujuh *dataset Twitter* Indonesia yang mencakup analisis sentimen, klasifikasi emosi, deteksi ujaran kebencian, dan pengenalan entitas nama (NER). Dari hasil percobaan, *IndoBERTweet* secara konsisten lebih unggul dibanding model lain seperti *MBERT*, *MalayBERT*, maupun *IndoBERT* standar. Yang menarik, ketika digunakan strategi *domain-adaptive pretraining* dengan rata-rata *embedding subword*, performanya bisa melampaui model yang dilatih dari awal, sekaligus menghemat waktu hingga lima kali lebih cepat. Hal ini menunjukkan bahwa *IndoBERTweet* tidak hanya berperan sebagai model bahasa untuk media sosial

Indonesia, tetapi juga menawarkan cara efisien dalam menyesuaikan model bahasa ke domain tertentu (Koto *et al*., 2021).

2.10. Analisis Sentimen

Analisis sentimen merupakan salah satu sub-bidang dalam penambangan teks (*text mining*) yang memiliki fokus utama pada proses klasifikasi dokumen. Metode ini memanfaatkan berbagai teknik pemrosesan bahasa alami (NLP) untuk mengambil dan mengenali pendapat, perasaan, dan penilaian yang ditulis seseorang mengenai suatu topik tertentu (Idris I *et al*., 2023). Analisis sentimen berfungsi mengenali arah dari pernyataan dalam teks. Arah pernyataan tersebut dapat mendukung, menolak, atau netral, yang ditentukan dari makna yang diutarakan oleh seseorang (Rivaldi *et al*., 2024).

Analisis sentimen memiliki peran penting untuk memahami persepsi publik atau opini pengguna terhadap suatu topik, produk, layanan, organisasi, maupun kebijakan (Tan *et al*., 2023). Selain itu, analisis sentimen juga dapat membantu pengambilan keputusan oleh perusahaan, institusi pemerintahan, atau organisasi lainya berdasarkan feedback dari masyarakat (Firdaus *et al*., 2024). Analisis ini juga berfungsi untuk menilai tingkat kepuasan, reputasi, serta kepercayaan publik secara objektif. Dengan demikian, analisis sentimen menjadi salah satu pendekatan yang relevan dalam memahami dinamika opini publik di era digital. Analisis ini juga berfungsi untuk menilai kepuasan pengguna serta menggambarkan reputasi suatu entitas secara umum. Dengan demikian, analisis sentimen dapat dipandang sebagai pendekatan yang relevan dalam memahami dinamika opini publik di era digital.

Analisis sentimen memiliki beberapa tantangan utama. Salah satunya adalah kebutuhan dataset yang baik, lengkap, dan beragam bahasa agar hasil penelitian lebih akurat. Tantangan lain muncul dari sulitnya mendeteksi emosi tersembunyi seperti sarkasme dan ironi, serta penggunaan bahasa gaul, singkatan, dan *emoji* yang sering muncul di media sosial. Selain itu, hasil analisis sering bergantung pada konteks, sehingga kata

yang sama bisa bermakna positif atau negatif tergantung situasinya. Masalah lain adalah model yang dilatih di satu bidang atau bahasa biasanya tidak bekerja baik di bidang atau bahasa lain (Mao *et al*., 2024).

2.11. *Emojional*

Emojional adalah pendekatan untuk membuat *emoji embedding* yang dirancang agar bisa menangkap emosi dan nuansa penggunaan *emoji* dengan lebih baik. Pendekatan ini muncul karena model-model *emoji* sebelumnya sering kesulitan memahami beragam emosi yang tersirat dalam *emoji* modern. Berbeda dengan model lain yang hanya mengandalkan dataset *Twitter* besar atau deskripsi *Unicode* yang terbatas, *Emojional* melihat *emoji* sebagai alat yang kaya informasi emosional, sehingga bisa menghasilkan *embedding* yang mencerminkan berbagai perasaan dan sensasi, tidak hanya satu jenis emosi (Barry *et al*., 2022).

Pendekatan ini dikembangkan karena *emoji* sangat penting dalam komunikasi media sosial, tetapi sering diabaikan dalam pemrosesan bahasa alami. Masalah utama yang ingin diatasi antara lain makna *emoji* yang terus berubah, misalnya *emoji* tinju yang bisa mendapatkan makna politik, dan penambahan *emoji* baru setiap tahun yang membuat model lama cepat usang. Selain itu, model sebelumnya seperti *emoji2vec*, yang hanya menggunakan deskripsi *Unicode*, sering gagal menangkap nuansa emosional, dan beberapa pendekatan berbasis kategori emosi terlalu sederhana sehingga tidak bisa menggambarkan beragam emosi sebuah *emoji* (Barry *et al*., 2022).

Keunikan utama dari *Emojional* terletak pada tiga hal besar. Pertama, *multifaceted emotional representation*, di mana model ini melihat *emoji* bukan hanya sebagai simbol visual, tetapi juga sebagai alat ekspresi emosi (*tonal device*) yang mampu membawa berbagai nuansa perasaan seperti marah, lucu, kecewa, hingga sarkastik secara bersamaan. Kedua, pendekatan *emotion dictionary-based*, yaitu model *Emojional* dilatih menggunakan deskripsi dari situs kamus *emoji* seperti *Emojipedia* dan *Emoji.Wiki*, bukan hanya dari *Unicode description*. Karena berbasis kamus, model ini tidak

memerlukan *dataset* berukuran besar dan lebih mudah diperbarui ketika muncul *emoji* baru atau ketika makna *emoji* berubah (Barry *et al*., 2022).

Ketiga, *Emojional* menerapkan *three-layered representation*, yaitu pendekatan yang memahami *emoji* melalui tiga tingkat makna. Pertama, Visual *atau Physical Description*, yaitu pemahaman terhadap bentuk *emoji* secara langsung, seperti *emoji* ✊ yang digambarkan sebagai *raised fist*. Kedua, *Emotional Nuance*, yaitu interpretasi makna emosional yang lebih halus, misalnya menggambarkan perasaan *resistance* (perlawanan) atau *defiance* (keberanian). Ketiga, *Social* atau *Political Context*, yaitu konteks sosial yang lebih luas, seperti keterkaitan *emoji* tersebut dengan gerakan sosial tertentu, misalnya *Black Lives Matter* (Barry *et al*., 2022).

Secara teknis, *Emojional* menggunakan algoritma *logistic regression*, yaitu metode dalam *machine learning* yang digunakan untuk mengukur seberapa cocok dua data memiliki hubungan tertentu (misalnya *emoji* dengan deskripsinya). Dalam hal ini, model berusaha menghubungkan *emoji* dengan kalimat penjelasannya di ruang vektor berdimensi 300, menggunakan acuan dari *Google News word2vec embeddings* yaitu representasi kata yang sudah dilatih dari miliaran teks berita (Barry *et al*., 2022).

Langkah pertama adalah *mapping description to vectors* (pemetaan deskripsi ke vektor). Setiap *emoji* biasanya memiliki deskripsi berupa urutan kata, misalnya "*happy face with tears of joy*". Tiap kata diubah ke bentuk vektor menggunakan *word2vec*, lalu semua vektor kata dijumlahkan menjadi satu vektor representatif untuk deskripsi *emoji* tersebut:

$$v_j = \sum_{k=1}^{n} w_k \tag{5}$$

di mana v_j adalah vektor hasil dari deskripsi *emoji* ke-j, dan $w_k$adalah vektor kata ke-k dalam deskripsi tersebut. Dengan cara ini, deskripsi panjang bisa disederhanakan menjadi satu vektor angka yang mewakili maknanya (Barry *et al*., 2022).

Setiap *emoji* juga memiliki vektor tersendiri, disebut x_i. Hubungan antara *emoji* (x_i) dan deskripsinya (v_j) dihitung menggunakan *dot product*, yaitu operasi matematika untuk

mengukur kesamaan dua vektor. Semakin besar hasil *dot product*-nya, semakin mirip maknanya. Hasilnya lalu diubah menjadi probabilitas menggunakan *sigmoid function* (σ), yang mengubah angka menjadi nilai antara 0 sampai 1 menunjukkan kemungkinan kecocokan antara *emoji* dan deskripsinya:

$$p(i,j) = \sigma(x_i^T v_j) \quad (5)$$

Untuk mengukur seberapa baik prediksi model, digunakan *loss function* berbasis *logistic loss*. Fungsi ini membantu model memperbaiki diri setiap kali salah menebak, agar prediksinya semakin akurat. Rumusnya adalah:

$$L(i,j,y_{ij}) = -\log\left(\sigma\left(y_{ij}x_i^T v_j - \left(1 - y_{ij}\right)x_i^T v_j\right)\right) \quad (6)$$

di mana $y_{ij} = 1$ jika deskripsi *j* benar-benar cocok dengan *emoji i*, dan jika $y_{ij} = 0$ jika tidak. Jadi, jika model salah menilai hubungan antara *emoji* dan deskripsi, nilainya akan dikoreksi lewat fungsi ini agar mendekati target yang benar (Barry *et al*., 2022).

Agar model tidak hanya mengenali pasangan *emoji*-deskripsi yang benar, digunakan teknik *negative subsampling* yaitu menambahkan contoh "salah" atau "tidak cocok" secara sengaja dalam proses pelatihan. Misalnya, *emoji* ☺ mungkin dipasangkan dengan deskripsi "*crying loudly*" sebagai contoh negatif. Dengan begitu, model belajar membedakan mana pasangan yang benar dan mana yang salah. Dalam setiap batch pelatihan, digunakan 8 contoh data (4 positif dan 4 negatif), sehingga model menjadi lebih seimbang dan hasil *embedding*-nya lebih akurat (Barry *et al*., 2022).

Dari segi performa, *Emojional* menunjukkan hasil yang sangat baik secara kuantitatif dan kualitatif. Dalam pengujian *sentiment analysis* menggunakan algoritma *Random Forests* dan *Linear SVM* , kombinasi *embeddings* dari *Emojional* dan *GloVe* embeddings menghasilkan akurasi tertinggi dibandingkan model lain seperti *emoji2vec*, *Sense_des*, dan *Face Emojis*. Model ini menunjukkan peningkatan akurasi sebesar 14.7% pada *Random Forests* dan 17.9% pada *Linear SVM* untuk 90% *emoji* yang paling sering digunakan. Artinya, memahami emosi dalam *emoji* dapat meningkatkan kinerja sistem *NLP* secara signifikan, terutama dalam analisis emosi dan opini di media sosial (Barry *et al*., 2022).

2.12. *Emoji Sentiment Ranking*

Emoji Sentiment Ranking merupakan leksikon sentimen *emoji* pertama yang dikembangkan sebagai sumber daya dalam analisis sentimen otomatis, khususnya untuk bahasa-bahasa di wilayah Eropa. Leksikon ini dibangun melalui analisis terhadap lebih dari 1,6 juta *tweet* dalam 13 bahasa Eropa, yang telah diberi label sentimen (positif, netral, atau negatif) oleh 83 penilai manusia. Nilai sentimen setiap *emoji* ditentukan berdasarkan kecenderungan sentimen dari *tweet* yang mengandung *emoji* tersebut, sehingga mencerminkan penggunaan *emoji* dalam konteks sebenarnya. Hasil penelitian menunjukkan bahwa sebagian besar *emoji*, terutama yang paling sering digunakan, memiliki kecenderungan positif. Selain itu, tidak ditemukan perbedaan signifikan pada peringkat sentimen *emoji* di antara ketigabelas bahasa yang dianalisis, sehingga *Emoji Sentiment Ranking* dinilai sebagai sumber daya yang bersifat independen terhadap bahasa (Novak *et al*., 2015).

Leksikon ini mencakup 751 *emoji* yang paling sering digunakan pada berbagai platform media sosial. Setiap *emoji* diberikan skor sentimen dengan rentang nilai −1 hingga +1, di mana nilai −1 menunjukkan sentimen paling negatif dan +1 menunjukkan sentimen paling positif. Skor tersebut dihitung berdasarkan probabilitas kemunculan *emoji* dalam *tweet* yang telah dilabeli negatif, netral, dan positif, sehingga menghasilkan ukuran kuantitatif terhadap kecenderungan emosional dari masing-masing *emoji*. Gambar 7 menunjukkan sebagian hasil dari *Emoji Sentiment Ranking* v1.0, yang berisi data mengenai kode *Unicode*, frekuensi kemunculan, serta skor sentimen negatif, netral, dan positif untuk setiap *emoji*. Nilai ini dihitung berdasarkan distribusi kemunculan *emoji* dalam *tweet* berlabel sentimen, dengan rentang skor −1 hingga +1 (Novak *et al*., 2015).

Emoji Sentiment Ranking v1.0

Char	Image [twemoji]	Unicode codepoint	Occurrences [5...max]	Position [0...1]	Neg [0...1]	Neut [0...1]	Pos [0...1]	Sentiment score [-1...+1]	Sentiment bar (c.i. 95%)	Unicode name	Unicode block
😂	😂	0x1f602	14622	0.805	0.247	0.285	0.468	0.221		FACE WITH TEARS OF JOY	Emoticons
❤	❤	0x2764	8050	0.747	0.044	0.166	0.790	0.746		HEAVY BLACK HEART	Dingbats
♥	♥	0x2665	7144	0.754	0.035	0.272	0.693	0.657		BLACK HEART SUIT	Miscellaneous Symbols
😍	😍	0x1f60d	6359	0.765	0.052	0.219	0.729	0.678		SMILING FACE WITH HEART-SHAPED EYES	Emoticons
😭	😭	0x1f62d	5526	0.803	0.436	0.220	0.343	-0.093		LOUDLY CRYING FACE	Emoticons
😘	😘	0x1f618	3648	0.854	0.053	0.193	0.754	0.701		FACE THROWING A KISS	Emoticons
😊	😊	0x1f60a	3186	0.813	0.060	0.237	0.704	0.644		SMILING FACE WITH SMILING EYES	Emoticons
👌	👌	0x1f44c	2925	0.805	0.094	0.249	0.657	0.563		OK HAND SIGN	Miscellaneous Symbols and Pictographs

Gambar 7. *Emoji Sentiment Ranking*
(Novak *et al*., 2015)

2.13. *Multidimensional Lexicon of Emojis*

Multidimensional Lexicon of Emojis (MLE) dikembangkan oleh Godard *et al*. dimana leksikon ini dibuat karena kebanyakan leksikon sebelumnya hanya menilai *emoji* berdasarkan dua hal saja, yaitu positif atau negatif. Dengan adanya *MLE*, analisis emosi bisa dilakukan dengan lebih lengkap dan mendalam karena mencakup sepuluh dimensi emosional. Delapan di antaranya menggambarkan emosi dasar, yaitu marah, antisipasi, jijik, takut, senang, sedih, terkejut, dan percaya, sedangkan dua lainnya mewakili sentimen umum, yaitu positif dan negatif (Godard & Holtzman, 2022). Gambar 8 menampilkan hasil pemetaan emosi dari *Multidimensional Lexicon of Emojis* (MLE) dalam bentuk tabel.

1	id	name	emoji	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	negative	positive
2	1	grinning face		0,15866736	0,55537039	0,11418585	0,19224357	0,56199283	0,18611749	0,28794155	0,59686768	0,35613543	0,97636319
3	2	grinning face with big eyes		0,17299551	0,58190243	0,13624115	0,21999113	0,60385499	0,19768102	0,30438643	0,59611651	0,38554416	0,99173507
4	3	grinning face with smiling eyes		0,17503777	0,51799942	0,13951132	0,18960909	0,51011252	0,18067127	0,24653779	0,52831404	0,38608701	0,88542767
5	4	beaming face with smiling eyes		0,17478679	0,5434456	0,12232378	0,20183014	0,53931657	0,18292927	0,28057811	0,55169181	0,37122959	0,9145014
6	5	grinning squinting face		0,23090421	0,46145731	0,16224326	0,23554034	0,43999789	0,21251874	0,24807827	0,486059	0,46075582	0,79962722
7	6	grinning face with sweat		0,22487112	0,46146809	0,17093543	0,24315107	0,39525954	0,22989622	0,2309843	0,46412586	0,46648252	0,76491964
8	7	rolling on the floor laughing		0,25952515	0,34433099	0,20563156	0,24023322	0,3221275	0,22797909	0,17949416	0,38736642	0,50967977	0,63570576

Gambar 8. *Multidimensional Lexicon of Emojis*
(Godard & Holtzman, 2022).

MLE berisi 359 *emoji* yang paling sering digunakan dalam komunikasi digital. *Emoji-emoji* tersebut dipilih berdasarkan data dari lebih dari tiga juta *tweet* yang dikumpulkan selama 17 bulan. Dalam proses pembuatannya, *MLE* menggunakan dasar dari *NRC Emotion Lexicon*, yaitu leksikon emosi yang sudah ada sebelumnya. Peneliti mengumpulkan tweet yang berisi *emoji*, lalu menganalisis teks di dalamnya dengan *NRC Emotion Lexicon* untuk menghitung skor setiap emosi. Nilai rata-rata dari skor tersebut digunakan untuk menentukan emosi yang paling sering muncul pada tiap *emoji*, sehingga hasilnya mencerminkan penggunaan *emoji* secara alami di media sosial.

Tahap selanjutnya adalah validasi hasil, yang dilakukan dengan membandingkan skor otomatis *MLE* dengan penilaian dari lebih dari 2.200 orang. Hasilnya menunjukkan bahwa sebagian besar emosi, terutama kategori senang dan negatif, memiliki tingkat kesesuaian yang tinggi. Hal ini menunjukkan bahwa *MLE* cukup akurat dan dapat

digunakan sebagai sumber tambahan dalam penelitian analisis sentimen yang melibatkan *emoji.*

2.14. Media Sosial

Media sosial dapat dipahami sebagai platform digital yang memungkinkan individu untuk berbagi informasi, ide, serta pengalaman secara luas. Selain sebagai sarana berbagi konten, media sosial juga berfungsi sebagai ruang interaksi dan komunikasi dalam kehidupan sehari-hari. Perkembangan teknologi membuat media sosial semakin mudah diakses dan digunakan oleh berbagai lapisan masyarakat. Pada akhirnya, media sosial berkembang menjadi salah satu media utama dalam menjalin hubungan dan berkomunikasi secara daring (Rahmau'dina *et al*., 2024).

Di Indonesia, sekitar 130 juta orang atau hampir setengah dari jumlah penduduk aktif menggunakan media sosial (Alfifa *et al*., 2025). Di antara berbagai platform, *Instagram* dan *Tiktok* menjadi yang paling banyak diminati masyarakat. Khususnya, *Tiktok* dengan jumlah pengguna aktif lebih dari 10 juta telah menjadikannya salah satu aplikasi berbagi video terpopuler (Alfifa *et al*., 2025),(Alfiah Zulqornain & Pandu Adikara, 2021)

Gambar 9. Logo Media Sosial *Facebook*, *Instagram*, *Tiktok.*

Instagram dan *Tiktok* banyak dimanfaatkan sebagai media komunikasi publik yang strategis. Di Sumatera Barat, misalnya, calon gubernur menggunakan *Instagram* untuk membangun citra politik melalui foto, video pendek, dan hashtag yang meningkatkan interaksi dengan masyarakat(Ahmad Suryadi & Jeni Mona Melisa, 2025). *Instagram*

juga terbukti berpengaruh dalam meningkatkan pengetahuan serta perilaku masyarakat, seperti dalam kampanye pengelolaan sampah plastik di Medan Tuntungan (Sihombing *et al*., 2024). Media sosial sebagai keseluruhan memainkan peran penting dalam pembentukan opini publik, menyebarkan narasi, memperkuat kampanye, mempercepat penyebaran informasi atau mis-informasi, memicu polarisasi melalui konten visual atau provokatif (Murodi & Setiyoko, 2024).

Dalam analisis sentimen di media sosial, peneliti sering menghadapi masalah data yang sangat besar, tidak seimbang, dan penuh dengan informasi tidak relevan. Data yang berantakan ini membuat proses pengolahan menjadi lebih sulit. Selain itu, model yang dibuat di satu bidang sering kali tidak bisa langsung bekerja dengan baik di bidang lain. Tantangan lain adalah banyak penelitian hanya fokus pada akurasi, padahal faktor praktis seperti kecepatan proses juga sangat penting (Xu *et al*., 2022).

2.15. *Feature Level Fusion* (Fusi Tingkat Fitur)

Feature-level fusion adalah metode yang menggabungkan fitur dari berbagai sumber dengan menyatukannya menjadi satu vektor fitur yang lebih besar. Pendekatan ini mempermudah integrasi informasi dari berbagai sumber sekaligus, sehingga dapat meningkatkan akurasi prediksi, dan karena itu sering digunakan dalam penelitian. Namun, metode ini memiliki beberapa tantangan, seperti adanya fitur yang redundan atau *noisy*, serta masalah *curse of dimensionality* ketika jumlah fitur terlalu banyak dan skala antar fitur berbeda. Untuk mengatasi hal ini, biasanya diterapkan pengurangan dimensi menggunakan teknik seperti *Principal Component Analysis* (PCA) dan normalisasi fitur dengan metode *min-max, z-score*, agar semua fitur berada pada skala yang sebanding (Al-Azani & El-Alfy, 2021).

Penggabungan fitur dilakukan dengan cara menyatukan fitur teks dan fitur *emoji* (Al-Azani & El-Alfy, 2021). Fitur teks merupakan representasi numerik dari kata-kata dalam teks, sementara fitur *emoji* dihitung berdasarkan frekuensi kemunculan *emoji*. Artinya, setiap *emoji* diberi nilai sesuai seberapa sering muncul dalam teks, kemudian

nilai tersebut dimasukkan ke dalam vektor fitur. Secara matematis, misalkan vektor fitur teks dinyatakan sebagai :

$$F = \{f_1, f_2, \dots., f_n\}, F \in \mathrm{R}^n , \quad (7)$$

dan vektor fitur *emoji* dinyatakan sebagai :

$$E = \{e_1, e_2, \dots., e_m\}, E \in \mathrm{R}^m , \quad (8)$$

Penggabungannya menghasilkan vektor baru :

$$C = \{f_1, f_2, \dots., f_n , e_1, e_2, \dots., e_m\}, F \in \mathrm{R}^k , \mathrm{k} = \mathrm{n+m} \quad (9)$$

2.16. *l2 Normalization*

L2 norm merupakan metode normalisasi yang digunakan untuk menyetarakan skala pada vektor numerik dengan cara membagi setiap elemen vektor dengan panjang total vektor tersebut. Panjang vektor dihitung menggunakan akar kuadrat dari jumlah kuadrat seluruh nilai dalam vektor. Melalui proses ini, nilai yang dihasilkan berada dalam rentang 0–1 dan vektor akhir memiliki magnitudo yang sama, yaitu satu (Fikri & Sarno, 2019). Dengan normalisasi ini, perbandingan antar dokumen menjadi lebih konsisten, dan tidak dipengaruhi oleh variasi panjang teks (Kusuma *et al*., 2025). *Rumus L2 normalization* ditunjukkan pada persamaan 10 berikut:

$$v_{norm} = \frac{v}{||v||^2} = \frac{v}{\sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}} \quad (10)$$

.Keterangan :

$\boldsymbol{v}$ = nilai vektor asli

$||\boldsymbol{v}||^{\boldsymbol{2}}$ = Panjang vektor atau nilai *L2 norm* dihitung dengan cara mengambil akar kuadrat dari penjumlahan kuadrat seluruh elemen dalam vektor. Dengan kata lain, setiap nilai pada vektor dikuadratkan terlebih dahulu, kemudian dijumlahkan, dan hasilnya diakarkan untuk mendapatkan nilai *L2 norm* tersebut.

2.17. *Adaptive Synthetic Sampling Approach* (ADASYN)

Adaptive Synthetic Sampling Approach (ADASYN) merupakan metode *oversampling* yang dirancang untuk mengatasi ketidakseimbangan kelas dalam sebuah *dataset*. Prinsip dasar *oversampling* adalah menambah jumlah data pada kelas minoritas agar proporsinya mendekati kelas mayoritas. *ADASYN* melakukan hal tersebut dengan cara menghasilkan sampel sintetis khusus pada area kelas minoritas yang dianggap lebih sulit diprediksi, yaitu dengan memanfaatkan informasi tentang distribusi lokal di sekitar titik data tersebut (Nurhayati & Rahardi, 2025). Pertama ADASYN menghitung derajat kelas yang tidak seimbang dengan rumus :

$$d = \frac{m_s}{m_i} \tag{11}$$

Dimana pada persamaan 11, d adalah derajat kelas yang tidak seimbang m_s adalah jumlah kelas minoritas dan m_i adalah jumlah kelas mayoritas. Setelah menghitung derajat kelas yang tidak seimbang, dilanjutkan dengan menghitung jumlah kelas minoritas yang yang perlu dibuat dengan rumus :

$$G = (m_i - m_s) \text{ x } \beta \tag{12}$$

Dimana β adalah parameter yang digunakan untuk menentukan level *balance* , dimana $\beta = 1$ artinya *dataset* yang sepenuhnya *balance.* Untuk setiap data minoritas *ADASYN* mencari *K- nearest neighbors* menggunakan jarak *Euclidean* dalam ruang fitur dan mengitung rasio r_i yang didefinisikan sebagai :

$$r_i = \frac{\Delta_i}{K} \tag{13}$$

Dimana $\Delta_i$adalah jumlah *K-nearest neighbors* yang dimiliki kelas mayoritas. r_i berada di rentang 0 sampai 1. Kemudian, normalisasi r_i dengan membagi r_i dengan jumlah semua r_i. Hasil normalisasi adalah $\hat{i}$. Setelah itu, hitung data sintetis yang perlu dibuat untuk tiap kelas minoritas dengan rumus :

$$g_i = \mathrm{r}^{\mathrm{i}} x\ G \tag{14}$$

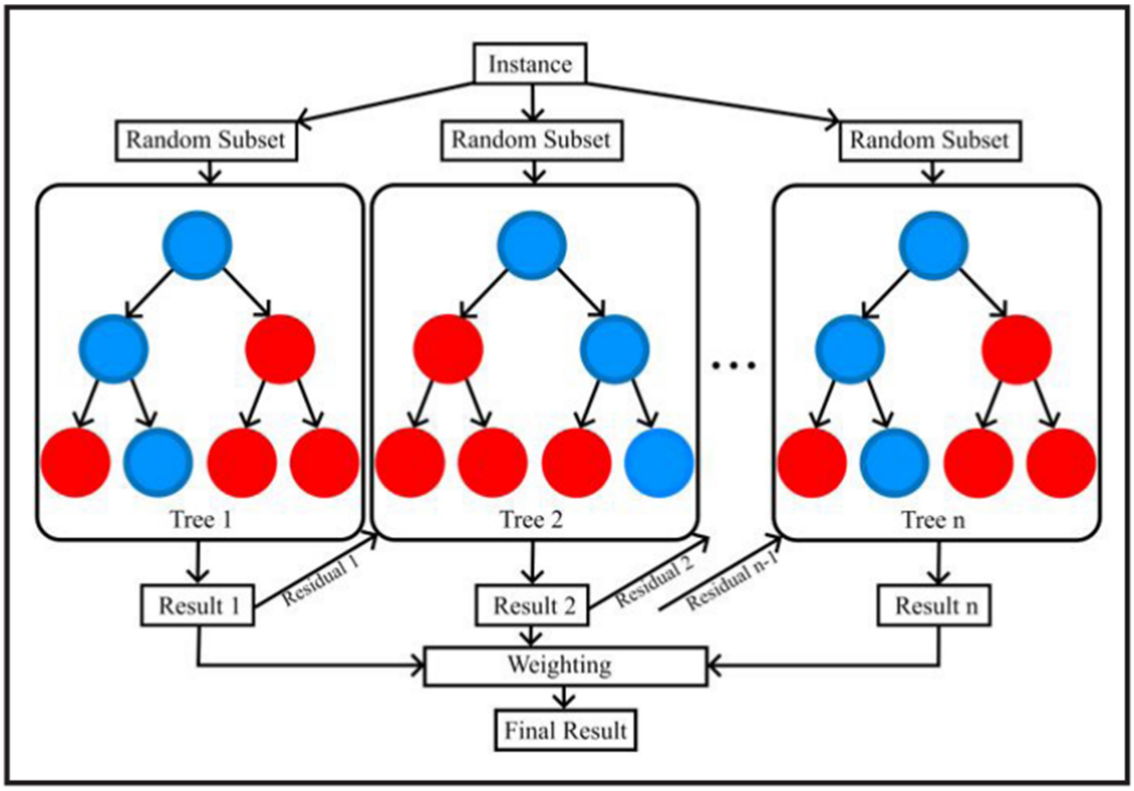
Dimana *G* adalah total jumlah data sintetis yang akan dibuat untuk kelas minoritas. Untuk setiap data pada kelas minoritas (x_i) buat data sintesis sebanyak g_i. Pada setiap prosesnya pilih satu data kelas minoritas (x_{zi}) terdekat dari x_i lalu buat data sintesis dengan rumus :

$$S_i = x_{\mathrm{i}} + (x_{zi} - x_i) \text{ x } \lambda \qquad (15)$$

Dimana $(x_{zi} - x_i)$ adalah nilai vektor selisih keduanya dan λ adalah angka acak 0 sampai 1 (He *et al*., 2008).

2.18. *Extreme Gradient Boosting* (XGBoost)

Gradient Boosting adalah metode *ensemble learning* yang bekerja dengan cara membangun model secara bertahap. Setiap model baru ditambahkan untuk memperbaiki kesalahan dari model sebelumnya. Umumnya, model dasar yang digunakan adalah pohon keputusan dengan kedalaman terbatas. Proses pembelajaran dilakukan secara iteratif , pertama dibentuk model awal sederhana, kemudian dihitung kesalahan atau residual, dan model baru dilatih untuk memprediksi kesalahan tersebut. Hasil akhirnya adalah kombinasi dari seluruh model kecil yang membentuk prediksi lebih akurat yang dapat dilihat pada gambar 10. Metode ini dikenal efektif meningkatkan akurasi, tetapi memiliki kelemahan seperti waktu pelatihan yang lama dan risiko *overfitting* jika parameter tidak diatur dengan baik (Bentéjac *et al*., 2019).



Gambar 10. Struktur *XGBoost*
(Oztornaci *et al*., 2024)

Extreme Gradient Boosting (XGBoost) merupakan pengembangan dari metode *gradient boosting* yang dirancang untuk lebih cepat, efisien, dan mampu menangani dataset besar. Inti perbedaan *XGBoost* adalah adanya regularisasi tambahan dan optimisasi teknis sehingga model yang dihasilkan lebih sederhana namun tetap akurat. Fungsi objektif *XGBoost* dituliskan sebagai:

$$CL_{xgb} = \sum_{i=1}^{N} L(y_i, F(X_i)) + \sum_{m=1}^{M} \Omega(h_m) \tag{16}$$

dengan fungsi regularisasi:

$$\Omega(h) = \gamma T + \frac{1}{2}\lambda || w ||^2 \tag{17}$$

di mana T adalah jumlah daun pada pohon, w adalah bobot pada setiap daun, γ mengatur seberapa besar penurunan *loss* yang dibutuhkan agar node dibelah, dan λ

adalah penalti kuadrat yang membuat bobot tidak terlalu besar. Dengan tambahan ini, *XGBoost* bisa mencegah *overfitting* sekaligus menghasilkan model yang lebih stabil (Bentéjac *et al*., 2019).

XGBoost memiliki sejumlah keunggulan teknis yang membuatnya lebih efisien. Pertama, algoritma ini menyimpan data dalam bentuk kolom yang sudah diurutkan terlebih dahulu. Tujuannya adalah agar proses pencarian titik pemisahan pada pohon keputusan lebih cepat, karena algoritma tidak perlu berulang kali mengurutkan data setiap kali membangun pemisahan. Kedua, *XGBoost* mendukung pemrosesan paralel, yaitu kemampuan membagi pekerjaan ke beberapa inti prosesor sekaligus. Dengan cara ini, waktu pelatihan model menjadi lebih singkat dibanding metode *boosting* biasa. Ketiga, *XGBoos*t menerapkan metode *approximate split finding*, yakni cara mencari pemisahan data dengan hanya mengevaluasi sebagian kandidat yang dianggap paling mewakili. Hal ini membuat perhitungan lebih ringan tanpa menurunkan kualitas model secara signifikan (Bentéjac *et al*., 2019).

Selain efisiensi, *XGBoost* juga unggul dalam menangani kualitas data. Algoritma ini bersifat *sparsity-aware*, artinya mampu mengelola data yang memiliki *missing values* tanpa perlu dilakukan pengisian manual (imputasi). Hal ini penting karena pada praktiknya, *dataset* sering tidak lengkap. Selain itu, *XGBoost* menggunakan teknik *row subsampling* dan *column subsampling*, yaitu mengambil sebagian data maupun sebagian fitur secara acak dalam proses pelatihan. Tujuan dari teknik ini bukan hanya mempercepat pembelajaran, tetapi juga mencegah *overfitting* dengan cara membuat model lebih bervariasi. Kinerja *XGBoost* kemudian diatur melalui parameter-parameter penting, seperti *learning rate* untuk menentukan seberapa besar langkah pembaruan model, *gamma* sebagai batas minimum agar suatu pemisahan dilakukan, *max depth* yang mengendalikan kedalaman pohon, *subsample* untuk menentukan proporsi data yang digunakan setiap iterasi, serta *colsample_bylevel* untuk mengatur jumlah fitur yang dipertimbangkan. Dengan kombinasi teknik dan pengaturan tersebut, *XGBoost* mampu memberikan hasil prediksi yang akurat dengan waktu pelatihan yang lebih singkat, sehingga lebih unggul dibandingkan *gradient boosting* konvensional maupun *random forest* (Bentéjac *et al*., 2019).

2.19. Metriks Evaluasi

Dalam mengevaluasi kinerja model yang dikembangkan, terdapat empat metrik yang umum digunakan, yaitu *accuracy*, *precision*, *recall*, *dan F1 Score* (Sathyanarayanan, 2024). *Accuracy* mengukur proporsi prediksi yang benar terhadap seluruh data, sebagaimana ditunjukkan pada persamaan 18 :

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \quad (18)$$

Metrik *precision* pada persamaan 19 digunakan untuk mengukur ketepatan model dalam memprediksi kelas positif, yang dirumuskan sebagai:

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (19)$$

Sementara itu, *recall* seperti pada persamaan 20 mengukur kemampuan model dalam menemukan seluruh data positif yang sebenarnya, dengan rumus:

$$Recall = \frac{TP}{TP + FN} \times 100\ \% \quad (20)$$

Untuk memperoleh keseimbangan antara *precision* dan *recall*, digunakan *F1 Score* sebagaimana pada persamaan 21:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \quad (21)$$

Keterangan :

- TP (*True Positive*) = Jumlah Prediksi benar dengan hasil positif
- TN (*True Negative*) = Jumlah Prediksi benar dengan hasil negatif
- FP (*False Positive*) = Prediksi positif dengan nilai asli negatif
- FN (*False Negative*) = Prediksi negatif dengan nilai asli positif.

III. METODE PENELITIAN

3.1. Tempat dan Waktu Penelitian

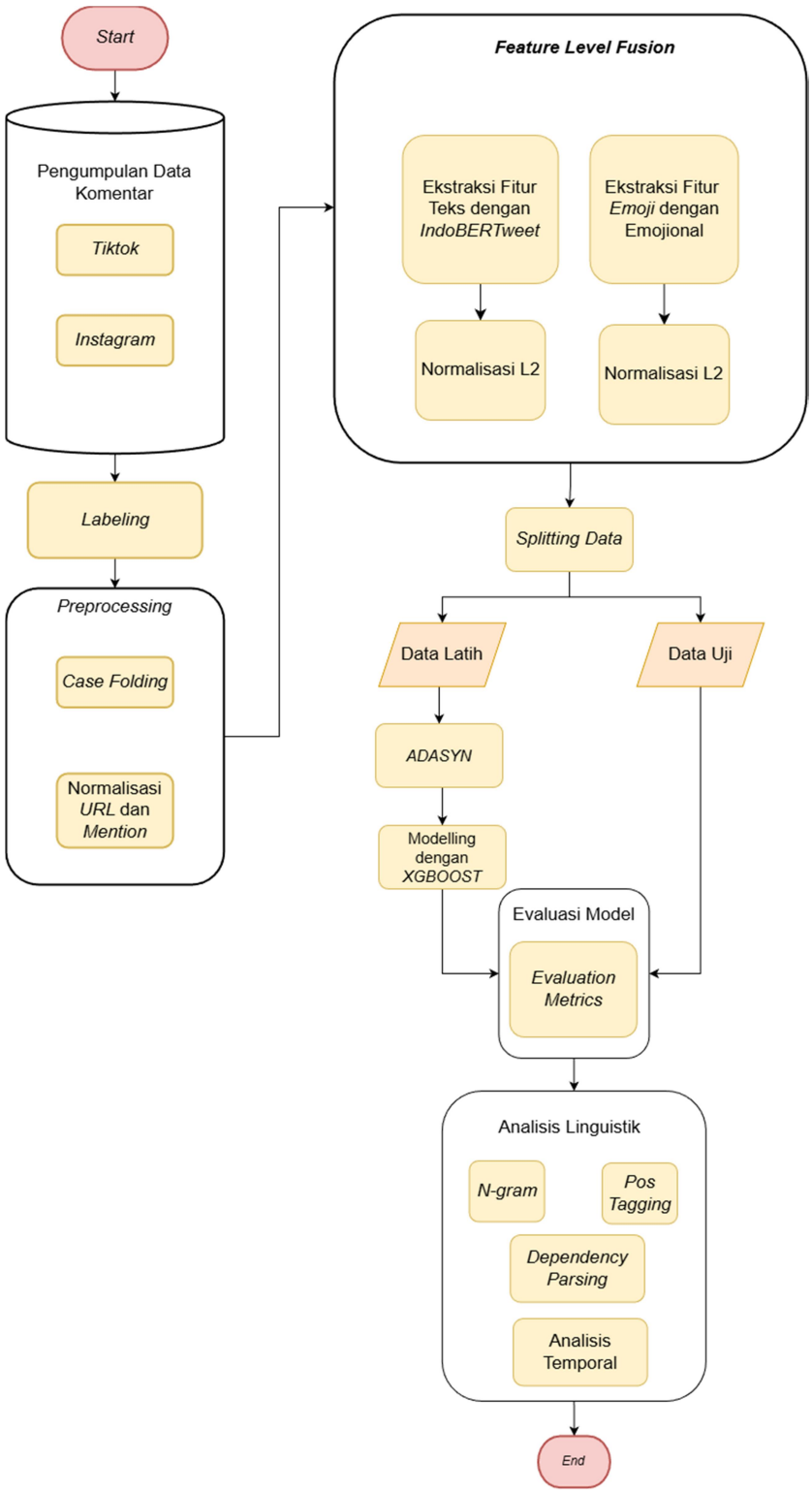
3.1.1. Tempat Penelitian

Penelitian ini dilakukan di Laboratorium Komputasi Dasar pada Program Studi Ilmu Komputer, Universitas Lampung. Lokasi ini dipilih karena menyediakan berbagai fasilitas yang mendukung pelaksanaan penelitian, seperti perangkat komputer dengan spesifikasi yang memadai, *software* untuk analisis, dan akses internet yang baik guna menunjang kegiatan pencarian literatur serta pengolahan data. Selain itu, lingkungan laboratorium yang tertib dan kondusif membantu peneliti dalam menjalankan proses penelitian secara optimal dan sesuai dengan rencana yang telah disusun.

3.1.2. Waktu Penelitian

Pelaksanaan penelitian direncanakan mulai dari bulan Agustus 2025 sampai dengan Desember 2025. Jadwal pelaksanaan penelitian secara rinci dapat dilihat pada Lampiran 1.

3.2. Tahapan Penelitian



Gambar 11. Tahapan Penelitian

3.2.1. Pengumpulan Data Komentar

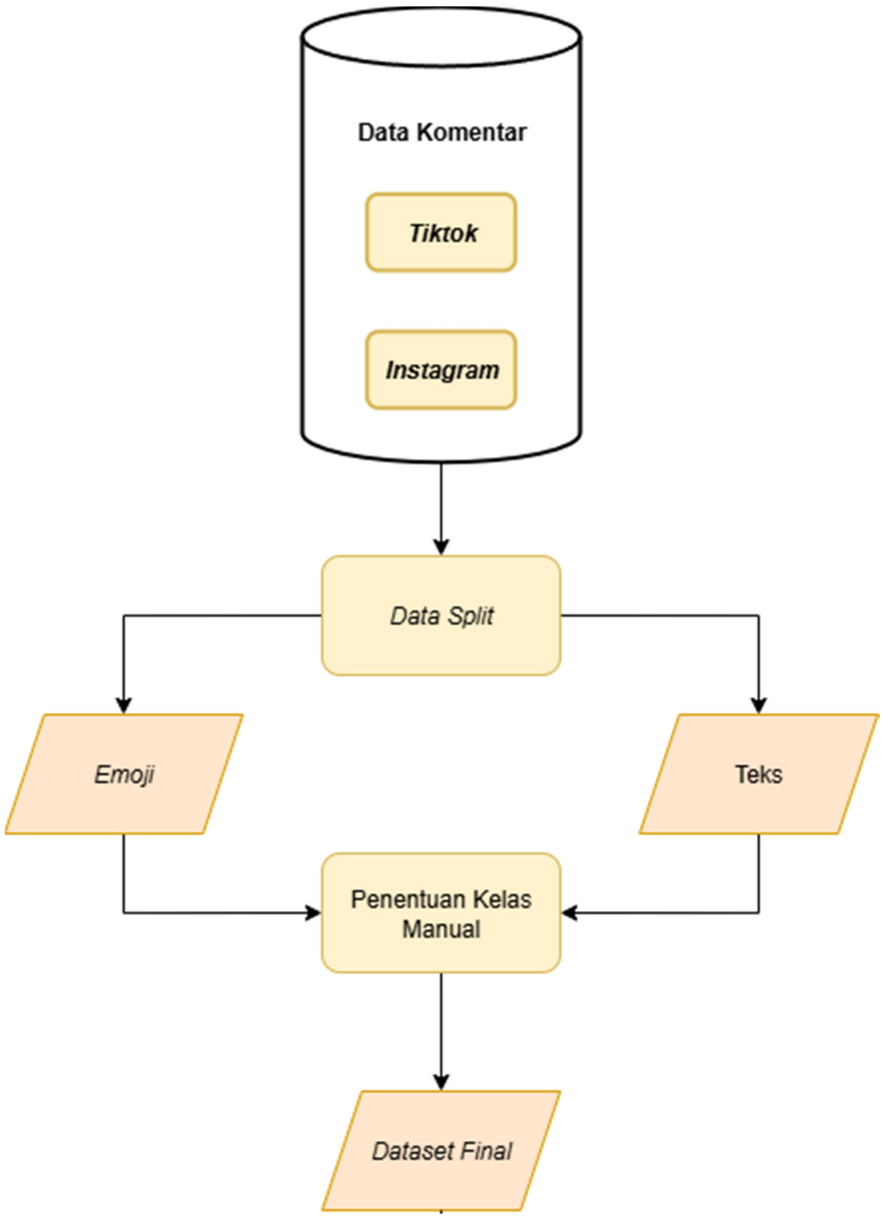
Dataset yang digunakan dalam penelitian ini berasal dari dua platform, yaitu *Tiktok* dengan 19.069 baris dan *Instagram* dengan 15.058 baris dengan total 34.127 baris. Setiap baris hanya memiliki satu kolom, yaitu *comment*, yang berisi teks komentar. Pengumpulan data komentar dilakukan menggunakan teknik *scraping* dengan bantuan ekstensi *Chrome Instant Data Scraper*. Tabel 2 menyajikan atribut yang terdapat pada dataset. Pada penelitian ini, dataset hanya terdiri dari satu atribut, yaitu comment, yang berisi teks komentar dari pengguna.

Tabel 2. *Dataset Features*

No.	Atribut	Keterangan
1.	*comment*	Tanggapan terhadap konten, seperti video, foto dalam bentuk kata, kalimat, maupun *emoji* pada kolom komentar.

3.2.2. *Labelling*

Proses penentuan label sentimen pada data komentar diawali dengan *dataset* komentar yang kemudian melalui tahap *data split*. Pada tahap ini, setiap komentar dipecah menjadi dua komponen utama, yaitu teks dan *emoji*. Pemisahan ini dilakukan bukan untuk pelatihan model secara langsung, melainkan untuk memudahkan proses pelabelan (*labelling*) pada tahap *pre-labelling*, sehingga masing-masing komponen dapat dianalisis secara independen. Seluruh proses dapat dilihat pada Gambar 12.



Gambar 12. Alur *Labelling*

Untuk komponen *emoji*, sentimen awal ditentukan dengan memanfaatkan *Emoji Sentiment Ranking* sebagai sumber utama (Novak *et al*., 2015). Namun, karena tidak semua *emoji* tercakup dalam leksikon tersebut, dilakukan pelengkapan data dengan mengacu pada leksikon dari penelitian Godard *et al*. yaitu *Multidimensional Lexicon of Emojis* guna menambah representasi *emoji* yang belum tersedia (Godard & Holtzman, 2022). Hasil analisis ini menghasilkan tiga kategori sentimen, yaitu negatif (0), netral (1), dan positif (2). Komponen teks dianalisis menggunakan model *IndoBERTweet*

untuk menghasilkan prediksi sentimen otomatis dengan tiga label yang sama, yaitu negatif (0), netral (1), dan positif (2). Rincian hasil sentimen dari komponen *emoji* dan teks ditampilkan pada Tabel 3. Penentuan label sentimen dilakukan dengan memilih skor tertinggi di antara kelas sentimen yang tersedia (0, 1, dan 2), baik berdasarkan leksikon *emoji* maupun hasil prediksi dari model *IndoBERTweet*. Skor tertinggi tersebut digunakan untuk menentukan label sementara pada masing-masing komponen sebelum peneliti menetapkan label akhir yang paling sesuai. Hasil prediksi dari kedua analisis tersebut kemudian diteruskan ke tahap krusial, yaitu penentuan label secara manual.

Tabel 3. Hasil Sentimen Awal dari Komponen *Emoji* dan Teks

No.	Komentar	Sentimen *Emoji*	Sentimen *Emoji* 2	Sentimen Teks	Label Akhir
1.	Emg smua nya itu ga masuk logika😂😂😂	Positif	-	Negatif	Negatif
2.	🔥🔥🔥 gubernurku	Netral	-	Postitif	Positif
3.	Salam dari tangsel👐👐👐👐	Positif	-	Netral	Netral
4	🌲🌲🌲 barakallah para pemimpin Lampung 🕌	Positif	Positif	Positif	Positif
5	sampah aja ga keurus, gimana mau ngurus gondola	-	-	Positif	Positif

Pada tahap manual ini, peneliti berperan sebagai penilai akhir yang mengevaluasi dan membandingkan saran sentimen dari kedua sumber otomatis dengan membaca dari

komentar asli untuk mengatasi ambiguitas serta menentukan label sentimen final yang paling akurat untuk setiap komentar. Setelah seluruh data dikoreksi dan dilabeli proses berakhir dengan terbentuknya *dataset* final yang siap digunakan pada tahap pelatihan model.

3.2.3. *Preprocessing*

Setelah dilakukan *labelling*, komentar pada *dataset* melalui tahap *preprocessing* untuk meningkatkan kualitas data. Tahap *preprocessing* yang diterapkan meliputi:

1. *Case folding* dan *lowercasing* untuk menyamakan seluruh huruf menjadi huruf kecil sehingga perbedaan huruf besar dan kecil tidak memengaruhi analisis.
2. Penggantian mention pengguna berupa *@nama* menjadi *@USER*
3. Penggantian *URL* menjadi kata *HTTPURL*

Langkah-langkah ini dilakukan untuk membersihkan data komentar sehingga lebih konsisten dan siap digunakan pada tahap analisis sentimen berikutnya.

3.2.4. *Feature Level Fusion*

Setelah tahap *preprocessing*, tahap selanjutnya dilakukan melalui beberapa langkah:

1. Pembentukan *Embedding* Teks
    - Setiap komentar yang telah melalui tahap *preprocessing* diproses satu per satu.
    - Teks dikonversi menjadi *token* menggunakan *tokenizer IndoBERTweet* agar dapat dipahami oleh model.
    - Model *IndoBERTweet* kemudian menghasilkan representasi vektor (*embedding*) dari *token* tersebut, khususnya menggunakan *token* [CLS] sebagai representasi keseluruhan makna komentar.
2. Pembentukan *Embedding Emoji*
    - Pada setiap komentar yang sama, sistem mendeteksi *emoji* yang terdapat di dalam teks.

- Setiap *emoji* yang ditemukan dikonversi menjadi representasi numerik menggunakan model *Emojional*, yang menangkap makna emosional dari *emoji* tersebut.
- Jika komentar tidak memiliki *emoji*, maka sistem menghasilkan vektor kosong (nol) sebagai pengganti agar tetap dapat digabung dengan *embedding* teks.

3. Normalisasi Menggunakan *L2 Normalization*

- Sebelum kedua *embedding* digabungkan, masing-masing *embedding* teks dan *emoji* terlebih dahulu dinormalisasi menggunakan teknik *L2 normalization*. Hal ini memastikan bahwa perbedaan skala antar *embedding* tidak mendominasi proses *fusion*, serta menjaga agar magnitude vektor berada pada rentang yang seragam. Dengan demikian, kedua *embedding* memiliki kontribusi yang lebih seimbang dalam proses penggabungan fitur.

4. Feature-Level Fusion

- Vektor hasil *embedding* teks dari *IndoBERTweet* digabungkan dengan *embedding emoji* dari model *Emojional*.
- Hasil penggabungan ini menghasilkan satu representasi fitur gabungan yang memuat informasi teks dan emosi, dan siap digunakan pada tahap klasifikasi.

3.2.5. *Splitting Data*

Setelah fitur gabungan diperoleh, *dataset* dibagi menjadi data pelatihan (*training set*) dan data pengujian (*test set*) untuk mendukung proses pembelajaran model. Pembagian dilakukan menggunakan metode *stratified sampling* agar distribusi label positif, netral, dan negatif tetap seimbang di kedua subset. Pembagian data pada penelitian ini menggunakan rasio 70%-30% dan 80%-20% sesuai dengan rekomendasi Kosheleva *et al*. yang menyatakan bahwa proporsi tersebut merupakan konfigurasi empiris terbaik (Kosheleva *et al*., 2018)

3.2.6. *Adaptive Synthetic Sampling* (ADASYN)

ADASYN diterapkan setelah proses pembagian data dan hanya dilakukan pada *data training*. Langkah ini memastikan bahwa *data testing* tetap merefleksikan distribusi asli. Pendekatan tersebut digunakan untuk mencegah *data leakage*, yaitu kondisi ketika informasi dari data *testing* tanpa sengaja masuk ke proses pelatihan sehingga membuat hasil evaluasi model menjadi tidak valid (Putra & Salam, 2025).

3.2.7. *Modelling*

Pada tahap ini, data latih yang telah diperoleh dari *ADASYN* sebelumnya dimanfaatkan untuk membangun model klasifikasi sentimen menggunakan algoritma *Extreme Gradient Boosting* (XGBoost). Algoritma ini dipilih karena kemampuannya dalam melakukan pembelajaran bertahap melalui pendekatan *gradient boosting decision tree*, di mana setiap iterasi menambahkan pohon keputusan baru untuk memperbaiki kesalahan dari model sebelumnya. *XGBoost* memiliki keunggulan dalam efisiensi komputasi, kemampuan menangani data berdimensi tinggi, serta penerapan mekanisme regularisasi yang efektif untuk mencegah terjadinya *overfitting* selama proses pelatihan (Bentéjac *et al*., 2019). Dengan kelebihan tersebut, *XGBoost* dinilai sesuai untuk mengolah representasi fitur gabungan yang dihasilkan dari *embedding IndoBERTweet* dan frekuensi *emoji*, sehingga diharapkan mampu memberikan hasil klasifikasi sentimen yang optimal.

Dalam pembangunan model, *XGBoost* memiliki sejumlah *hyperparameter* yang dapat disesuaikan untuk mengoptimalkan kinerja model. *Hyperparameter* ini mengatur aspek seperti kedalaman pohon keputusan, laju pembelajaran, jumlah pohon yang dibangun, serta tingkat regularisasi untuk mengendalikan kompleksitas model. Penyesuaian nilai-nilai tersebut dapat memengaruhi performa akhir model secara signifikan. Daftar beberapa *hyperparameter* penting beserta tipe, rentang nilai, dan deskripsi singkatnya ditunjukkan pada Tabel 4 berikut (Tarwidi *et al*., 2023).

Tabel 4. *Hyperparameter* Model *XGBOOST*

Hyperparameter	***Range***	***Default***	**Keterangan**
colsample_bytree	(0, 1]	1	Persentase fitur (kolom) yang digunakan untuk membangun setiap pohon keputusan.
learning_rate	[0, 1]	0.3	Ukuran langkah yang digunakan pada setiap iterasi ketika fungsi objektif dioptimalkan.
max_depth	[0, ∞)	6	Kedalaman maksimum setiap pohon, semakin besar nilainya maka model menjadi lebih kompleks dan berpotensi mengalam *overfitting.*
subsample	(0, 1]	1	Proporsi data yang diambil secara acak untuk membangun setiap pohon.
min_child_weight	[0, ∞)	1	Bobot minimum *instance* yang dibutuhkan agar *node child* dapat terbentuk.
n_estimators	[0, ∞)	100	Jumlah maksimum pohon yang dibangun oleh model selama proses *boosting.*
gamma	[0, ∞)	0	Pengurangan nilai *loss minimum* yang diperlukan untuk melakukan pemisahan

Hyperparameter	*Range*	*Default*	**Keterangan**
			node baru pada pohon.
alpha	[0, ∞)	0	Parameter regularisasi L1 yang berfungsi membuat model lebih konservatif dalam mempelajari fitur.
lambda	[0, ∞)	1	Parameter regularisasi L2 yang digunakan untuk mengontrol kompleksitas model agar tidak terjadi *overfitting.*

3.2.8. Evaluasi Model

Setelah model dilatih, tahap berikutnya adalah evaluasi kinerjanya untuk menilai kemampuan model dalam mengklasifikasikan komentar menjadi positif, netral, atau negatif. Evaluasi dilakukan menggunakan metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score* untuk mengukur kinerja model secara keseluruhan maupun per kelas, serta memastikan model bekerja dengan baik pada semua kategori.

3.3. Analisis Linguistik

Analisis linguistik dalam penelitian ini bertujuan untuk mengidentifikasi penyebab terbentuknya sentimen masyarakat terhadap kinerja Gubernur Lampung serta faktor-faktor yang menyebabkan perubahan sentimen dari bulan ke bulan. Untuk mencapai tujuan tersebut, penelitian ini menggunakan teknik *Natural Languange Processing* (NLP) seperti *N-gram*, *Part-of-Speech* (POS) *Tagging*, dan *Dependency parsing* guna

mengungkap pola isu, penggunaan bahasa, serta hubungan antar unsur kata dalam komentar masyarakat.

3.3.1. *N-gram*

Analisis *n-gram* digunakan untuk mengidentifikasi pola kemunculan kata dan sekumpulan kata yang sering muncul dalam komentar pengguna, untuk mengetahui hubungannya dengan komentar (Surendro *et al*., 2024). Dalam penelitian ini, *n-gram* dibuat dalam bentuk *unigram*, *bigram*, dan *trigram* untuk menangkap kata tunggal, pasangan kata berurutan, dan tiga kata yang berurutan. Sebelum analisis, dilakukan proses pembersihan teks dengan menghapus kata-kata yang tidak relevan menggunakan *stopword* yang dibuat secara manual. Selain itu, teks juga dinormalisasi menggunakan kamus *lexicon*(Aliyah Salsabila *et al*., 2018; Saputri *et al*., 2018). Selanjutnya, *n-gram* yang diperoleh dikelompokkan berdasarkan label sentimen, untuk masing-masing platform, *Instagram* dan *Tiktok*. Penggunaan *n-gram* digunakan untuk mengetahui isu yang menyebabkan sentimen positif, negatif, maupun netral antar platform *Instagram* dan *Tiktok* tiap bulannya maupun secara umum.

3.3.2. *Part of Speech* (POS) *Tagging*

*Part-of-Speech (*POS*) tagging* dilakukan untuk mengelompokkan setiap kata dalam komentar ke kategori tertentu, seperti kata benda, kata kerja, kata sifat, dan kategori (Kurniawan *et al*., 2022). Sebelum melakukan *Part-of-Speech (*POS*) tagging*, akan dilakukan normalisasi slang menggunakan kamus *lexicon* (Aliyah Salsabila *et al*., 2018; Saputri *et al*., 2018) tujuannya agar mengurangi variasi kata dengan bentuk yang sama. *Part of Speech* (POS) *Tagging* diimplementasikan menggunakan *library Stanza* Bahasa Indonesia, prosesnya mencakup tokenisasi, *multi-word token*, pemberian nilai *POS*, dan lematisasi. Setiap kata diberi label kategori *POS*, lalu dihitung berapa kali masing-masing kategori muncul. Frekuensi ini dianalisis tidak hanya berdasarkan label

sentimen komentar, tetapi juga dikelompokkan menurut platform yaitu *Instagram* dan *tiktok*.

Tujuan dari analisis ini adalah untuk memahami cara masyarakat menulis komentar dan gaya bahasanya. Misalnya, banyaknya kata sifat pada komentar negatif bisa menunjukkan penilaian atau ekspresi emosi, sementara kata kerja yang sering muncul dapat mengindikasikan tuntutan atau harapan terhadap suatu tindakan. Dengan begitu, *POS tagging* membantu melihat pola bahasa yang digunakan masyarakat.

3.3.3. *Dependency parsing*

Dependency parsing akan menggunakan *library Stanza* untuk memetakan struktur hubungan antar kata dalam setiap kalimat. Sebelum melakukan *dependency parsing* akan dilakukan normalisasi slang menggunakan kamus *lexicon* (Aliyah Salsabila *et al*., 2018; Saputri *et al*., 2018) tujuannya agar mengurangi variasi kata dengan bentuk yang sama. Dari hasil *dependency parsing*, akan diketahui *dependency relation* seperti *nsubj* (subjek), *obj* (objek), dan root (kata benda). *Dependency relation* ini kemudian akan dipetakan menjadi subjek–verba–objek (SVO). Hasil *analisis dependency parsing* kemudian digabungkan dengan label sentimen untuk mengetahui subjek, predikat, dan objek yang paling sering muncul pada komentar positif, netral, maupun negatif. Selanjutnya, hasil tersebut dikelompokkan berdasarkan platform, yaitu *Instagram* dan *Tiktok*. Dengan mengetahui subjek dapat diketahui pelaku atau pokok pembicaraan, dengan mengetahui verba dapat diketahui proses, perbuatan, atau keadaannya, dan dengan mengetahui objek dapat diketahui hal, perkara, atau orang yang menjadi fokus pembicaraan.

3.3.4. Analisis Temporal

Analisis temporal dilakukan untuk memantau perubahan distribusi sentimen dari waktu ke waktu. Pada tahap ini, setiap komentar yang telah dianalisis sentimennya dikelompokkan berdasarkan bulan unggahan dan platform media sosial, yaitu *Instagram*

dan *Tiktok*. Untuk setiap bulan, jumlah komentar beserta proporsi sentimen negatif, netral, dan positif dihitung. Hasil ini kemudian divisualisasikan dalam bentuk grafik atau histogram, sehingga dapat terlihat tren peningkatan atau penurunan sentimen dari bulan ke bulan. Selain itu, melalui analisis *n-gram*, dapat diidentifikasi kata atau frasa yang paling berkontribusi terhadap sentimen pada periode tertentu. Dengan pendekatan ini, dapat diketahui isu-isu yang menyebabkan sentimen positif maupun negatif di tiap bulannya.

V. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan penelitian yang dilakukan, didapatkan tiga kesimpulan sebagai berikut :

1. Hasil penelitian menunjukan masyarakat bahwa pada platform *Instagram*, sentimen positif mendominasi, diikuti sentimen negatif dan netral. Pada platform *Instagram* di bulan Februari tahun 2025, sentimen positif menunjukan banyak dukungan masyarakat terhadap awal kepemimpinan gubernur yang baru dilantik di bulan Febuari 2025, sementara sentimen negatif masih terbatas pada isu permintaan penghapusan uang komite dan keluhan jalan. Pada platform *Instagram* di bulan Maret 2025, sentimen positif menyoroti terkait perbaikan jalan dan program lampung mengajar, tetapi sentimen negatif menyoroti terkait jalan rusak dan banjir, hal ini menunjukkann masih terdapat perbaikan jalan yang belum merata. Pada platform *Instagram* di bulan April 2025 menunjukkan puncak sentimen positif yang fokus pada sektor olahraga yaitu sepak bola Lampung, pemutihan pajak, dan pada kebijakan pemutihan pajak, sedangkan sentimen negatif tinggi terkait belanja pegawai, ketidakadilan pemutihan pajak, dan kondisi jalan. Pada platform *Instagram* dari bulan Mei 2025 hingga Juni 2025, dukungan positif tetap pada perbaikan jalan dan stabilisasi harga harga singkong, sementara sentimen negatif muncul terkait pemotongan harga singkong, pemutihan pajak, jalan rusak, dan uang komite. Pada platform *Instagram* di bulan Juli 2025, sentimen positif fokus pada sepakbola lampung dan perbaikan fasilitas olahraga yaitu stadion sumpah pemuda, tetapi sentimen negatif tetap terkait uang komite, sekolah swasta, dan kondisi jalan. Pada platform *Instagram* di bulan Agustus 2025 dan September 2025, sentimen positif mengarah pada petani singkong dan sekolah rakyat, sementara sentimen negatif masih terkait jalan rusak dan harga singkong.

Di platform *Tiktok*, sentimen negatif lebih mendominasi, diikuti sentimen positif dan netral. Pada platform *Tiktok* di bulan Februari 2025, kritik utama terkait infrastruktur dan lingkungan, terutama banjir dan jalan rusak, sedangkan sentimen positif muncul pada dukungan terhadap kepempinan gubernur Lampung yang baru saja dilantik. Pada platform *Tiktok* di bulan Maret 2025, sentimen negatif tetap tinggi dengan keluhan jalan rusak parah dan harga singkong, sementara sentimen positif muncul pada perbaikan jalan. Pada platform *Tiktok* di bulan April 2025 menjadi puncak sentimen negatif karena ketidakpuasan terhadap program pemutihan pajak dan belanja pegawai yang dinilai tinggi, sedangkan sentimen positif muncul terkait program pemutihan pajak. Pada platform *Tiktok* di bulan bulan Mei 2025, sentimen negatif menyoroti harga singkong, praktik pemotongan harga oleh pabrik, dan jalan rusak, sentimen positif fokus pada penghapusan uang komite dan stabilisasi harga singkong. Pada platform *Tiktok* di bulan Juni 2025 menunjukkan sentimen negatif fokus terhadap pelaksanaan kebijakan penghapusan uang komite dan ketidakpuasan petani singkong terhadap kondisi harga singkong, sedangkan sentimen positif muncul pada hal yang sama yaitu penghapusan uang komite dan kondisi petani singkong. Pada platform *Tiktok* di bulan Juli 2025, sentimen negatif fokus pada nasib PPPK dan tenaga pendidik, sementara positif muncul pada apresiasi terhadap pembangunan fasilitas olahraga. Pada platform *Tiktok* di bulan Agustus 2025, kritik mengarah kepada nasib PPPK, harga singkong, dan jalan rusak, sementara sentimen positif lebih fokus pada pujian dan harapan terhadap gubernur Lampung kedepannya. Pada platform *Tiktok* di bulan September 2025, sentimen negatif masih terkait harga singkong, ketidakpuasan petani singkong, jalan rusak, praktik pemotongan harga oleh pabrik, dan mobil batu bara, sedangkan sentimen positif muncul pada larangan mobil batu bara, perbaikan jalan, dan petani singkong.

2. Hasil pengujian menunjukkan bahwa model yang memadukan *IndoBERTweet* dengan *Emojional* mencapai performa paling optimal pada pembagian data 70–30, dengan akurasi sebesar 89.247% dan F1-score sebesar 84.27%. Nilai ini lebih tinggi dibandingkan model yang hanya mengandalkan fitur teks, sehingga menegaskan bahwa penyertaan informasi dari *emoji* memberikan kontribusi nyata dalam meningkatkan kualitas prediksi. Konsistensi hasil juga terlihat pada seluruh

skenario pengujian, di mana model berbasis kombinasi teks dan *emoji* selalu menghasilkan nilai *accuracy* dan *F1-score* yang lebih baik daripada model tanpa fitur *emoji*. Temuan ini memperlihatkan bahwa *emoji* memiliki peran penting dalam menggambarkan emosi dan intensi pengguna pada komentar media sosial, dan pemanfaatannya melalui *embedding Emojional* membantu model menangkap konteks sentimen secara lebih mendalam.

3. Hasil analisis linguistik menunjukkan bahwa penyebab sentimen positif terhadap Gubernur Lampung di media sosial *Instagram* antara lain perbaikan jalan, pembangunan infrastruktur sepak bola dan kepedulian terhadap olahraga di provinsi Lampung, cara kerja dan responsivitas gubernur, program pemutihan pajak, stabilisasi harga singkong, pembebasan uang komite, dan program Lampung Mengajar. Di *Tiktok*, penyebab sentimen positif serupa, mencakup apresiasi terhadap perbaikan jalan, integritas dan karakter kepemimpinan, efektivitas kepemimpinan, program pemutihan pajak, stabilisasi harga singkong, pembebasan uang komite, serta kebijakan pelarangan truk batu bara. Penyebab sentimen negatif di *Instagram* berkaitan dengan kerusakan jalan, persebaran masalah infrastruktur di beberapa wilayah, ketidakpuasan terhadap pemutihan pajak yang dianggap tidak adil dan proses pembayaran pajak yang rumit, potongan harga singkong di tingkat pabrik, pungutan komite di sekolah yang masih terjadi, belanja pegawai yang dianggap terlalu tinggi, nasib pegawai honorer, sekolah swasta yang ingin mendapat pembebasan komite, dan masalah banjir. Sementara di *Tiktok*, penyebab sentimen negatif meliputi jalan rusak, potongan harga singkong dan tuntutan pengawasan pabrik, belum dilantiknya PPPK paruh waktu, praktik pungutan di sekolah, ketidakpuasan terhadap pemutihan pajak, mobil batu bara, masalah banjir, dan belanja pegawai yang dinilai mubazir. Penyebab sentimen netral di *Instagram* mencakup kebutuhan masyarakat akan kejelasan administrasi pemutihan pajak, pertanyaan terkait layanan Samsat keliling dan drive thru, kepastian mengenai kelayakan pemutihan pajak, serta pertanyaan apakah kebijakan pembebasan komite juga berlaku untuk sekolah swasta. Di *Tiktok*, penyebab sentimen netral menunjukkan fokus yang serupa dengan *Instagram*, yakni pertanyaan dan verifikasi terkait prosedur administrasi dan kepastian pelaksanaan kebijakan. Berdasarkan hasil *POS tagging* dan *dependency*

parsing platform *Instagram* terkait Gubernur Lampung, hasil *POS tagging* menunjukkan bahwa pada sentimen positif, kata benda merujuk pada Gubernur, pihak penerima manfaat, dan isu yang berhasil ditangani, kata kerja mengekspresikan dukungan dan apresiasi seperti "dukung" atau "lanjut", dan kata sifat menunjukkan optimisme dan penilaian baik terhadap kinerja seperti "mantap", "sehat", dan "keren", sedangkan hasil *dependency parsing* mengungkapkan bahwa subjek komentar umumnya adalah penerima manfaat kebijakan, kata kerja terkait apresiasi program pemerintah seperti pemutihan pajak, perbaikan jalan, dan stabilisasi harga, dan objek mencakup masyarakat, Gubernur, serta program yang dijalankan. Pada sentimen negatif, *POS tagging* menunjukkan kata benda merujuk pada pihak terdampak masalah dan isu konkret seperti kerusakan jalan, harga singkong, belanja pegawai, serta uang komite, kata kerja menekankan keluhan dan desakan perbaikan, dan kata sifat menyoroti kritik terhadap kondisi dan kinerja, sementara *dependency parsing* menampilkan subjek yang mewakili pihak terdampak langsung, kata kerja yang menekankan keluhan dan tuntutan perbaikan, dan objek yang mencakup Gubernur, kondisi infrastruktur, anggaran, dan program pemerintah. Untuk sentimen netral, *POS tagging* menunjukkan kata benda terkait prosedur administratif dan pemutihan pajak, kata kerja dan kata sifat bersifat informatif dan deskriptif tanpa muatan emosional, sedangkan *dependency parsing* mengungkapkan subjek terkait layanan Samsat dan dokumen kendaraan, kata kerja menekankan penyampaian informasi atau instruksi, dan objek fokus pada administrasi pajak, fasilitas publik, dan program pemerintah, termasuk pelaporan atau permintaan informasi terkait fasilitas dan pengelolaan lingkungan. Analisis *dependency parsing* dan *POS tagging* pada platform *TikTok* tentang Gubernur Lampung menunjukkan perbedaan pola bahasa menurut sentimen. Pada sentimen positif, *POS tagging* memperlihatkan kata benda terkait Gubernur, pihak penerima manfaat, dan isu yang berhasil ditangani, kata kerja mengekspresikan dukungan dan apresiasi, sedangkan kata sifat menandakan optimisme terhadap kinerja, sementara *dependency parsing* menampilkan subjek sebagai penerima manfaat kebijakan, kata kerja menyoroti apresiasi terhadap program seperti pemutihan pajak dan perbaikan jalan, dan objek mencakup masyarakat, Gubernur, serta program yang dijalankan. Pada sentimen

negatif, *POS tagging* menunjukkan kata benda terkait pihak terdampak masalah dan isu konkret, kata kerja menekankan keluhan, dan kata sifat menyoroti kritik, sedangkan *dependency parsing* menampilkan subjek terdampak langsung, kata kerja menekankan desakan perbaikan, dan objek mencakup Gubernur, infrastruktur, anggaran, serta program pemerintah. Untuk sentimen netral, *POS tagging* menunjukkan kata benda terkait prosedur administratif dan pemutihan pajak, kata kerja dan kata sifat bersifat informatif, sedangkan *dependency parsing* menampilkan subjek terkait layanan Samsat, kata kerja menekankan penyampaian informasi, dan objek fokus pada administrasi pajak, fasilitas publik, dan program pemerintah.

5.2 Saran

Adapun saran dari penelitian yang telah dilakukan yaitu :

1. Penelitian selanjutnya dapat memperluas sumber data ke platform lain seperti *Twitter* atau *YouTube* untuk memberikan gambaran opini publik yang lebih komprehensif.

2. Penelitian selanjutnya dapat mengembangkan analisis sentimen berbasis teks dan *emoji* dengan teknik fusi lain, seperti *score level fusion atau decision level fusion*, yang berpotensi meningkatkan akurasi model, khususnya dalam menangani komentar yang kaya akan *emoji*.

DAFTAR PUSTAKA

Ahmad Suryadi, & Jeni Mona Melisa. (2025). Instagram and Political Power: How Social Media Enhances the Self-Image of Political Actors. *Al-Jamahiria : Jurnal Komunikasi Dan Dakwah Islam*, *3*(1), 57–69. https://doi.org/10.30983/al-jamahiria.v3i1.9925

Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., Kurniawan, K., Moeljadi, D., Prasojo, R. E., Baldwin, T., Lau, J. H., & Ruder, S. (2022). One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1*, 7226–7249. https://doi.org/10.18653/v1/2022.acl-long.500

Al-Azani, S., & El-Alfy, E. S. M. (2021). Early and Late Fusion of Emojis and Text to Enhance Opinion Mining. *IEEE Access*, *9*, 1–15. https://doi.org/10.1109/ACCESS.2021.3108502

Alexander, S. G., Ananto, A. T., Adhitya, I. P., Mangku, P., & Liano, B. (2023). Analisis Sentimen Opini Masyarakat Indonesia Terhadap Konten Deepfake Tokoh Publik. *KAKIFIKOM (Kumpulan Artikel Karya Ilmiah Fakultas Ilmu Komputer)*, *05*(02), 95–102. https://ejournal.ust.ac.id/index.php/KAKIFIKOM/article/view/3073%0Ahttps://ejournal.ust.ac.id/index.php/KAKIFIKOM/article/view/3073/2545

Alfiah Zulqornain, J., & Pandu Adikara, P. (2021). Analisis Sentimen Tanggapan Masyarakat Aplikasi Tiktok Menggunakan Metode Naïve Bayes dan Categorial Propotional Difference (CPD). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *5*(7), 2886–2890. http://j-ptiik.ub.ac.id

Alfifa, A. N., Fahira, V., Batubara, R., Sakinah, K., & Yoes, R. A. (2025). Media Sosial dan Pembentukan Opini Publik (Analisis Studi Kasus Echo Chamber Pada Interaksi Komentar di Akun Instagram @ Turnbackhoaxid Dalam Konteks Post – Truth). *Socius: Jurnal Penelitian Ilmu-Ilmu Sosial*, *2*(6), 162–169.

https://doi.org/doi.org/10.5281/zenodo.14709277

Aliyah Salsabila, N., Ardhito Winatmoko, Y., Akbar Septiandri, A., & Jamal, A. (2018). Colloquial Indonesian Lexicon. *2018 International Conference on Asian Language Processing (IALP)*, 226–229. https://doi.org/10.1109/IALP.2018.8629151

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. In *Journal of Big Data* (Vol. 8, Issue 1). Springer International Publishing. https://doi.org/10.1186/s40537-021-00444-8

Amien, M. (2023). Sejarah dan Perkembangan Teknik Natural Language Processing (NLP) Bahasa Indonesia: Tinjauan tentang sejarah, perkembangan teknologi, dan aplikasi NLP dalam bahasa Indonesia. *ELANG:Journal of Interdisciplinary Research*, *1*(1), 99–105. https://doi.org/https://doi.org/10.32664/elang.v1i01

Arora, M., Kanjilal, U., & Varshney, D. (2016). Evaluation of information retrieval: precision and recall. *International Journal of Indian Culture and Business Management*, *12*(2), 224. https://doi.org/10.1504/ijicbm.2016.074482

Astuti, K. C., Firmansyah, A., & Riyadi, A. (2024). Implementasi Text Mining Untuk Analisis Sentimen Masyarakat Terhadap Ulasan Aplikasi Digital Korlantas Polri pada Google Play Store. *REMIK: Riset Dan E-Jurnal Manajemen Informatika Komputer*, *8*(1), 383–394.

Azahra, N. M., & Setiawan, E. B. (2023). Sentence-Level Granularity Oriented Sentiment Analysis of Social Media Using Long Short-Term Memory (LSTM) and IndoBERTweet Method. *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika (JITEKI)*, *9*(1), 85–95. https://doi.org/10.26555/jiteki.v9i1.25765

Barry, E., Jameel, S., & Raza, H. (2022). Emojional: Emoji Embeddings. In T. Jansen, R. Jensen, N. Mac Parthaláin, & C.-M. Lin (Eds.), *Advances in Computational Intelligence Systems Contributions Presented at the 20th UK Workshop on Computational Intelligence, September 8-10, 2021, Aberystwyth, Wales, UK. Part of the Advances in Intelligent Systems and Computing book series (AISC, volume* (Vol. 1409, pp. 312–324). Springer. https://repository.essex.ac.uk/30781/

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). *A Comparative Analysis of XGBoost*. *November 2019*. https://doi.org/10.1007/s10462-020-09896-5

Cardova, A., & Hermawan, A. (2023). Implementasi Metode LSTM Untuk Mengklasifikasi Berita Palsu Pada PolitiFact. *Jurnal Fasilkom*, *13*(3), 471–479. https://doi.org/10.37859/jf.v13i3.6175

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Fikri, M., & Sarno, R. (2019). A comparative study of sentiment analysis using SVM and SentiWordNet. *Indonesian Journal of Electrical Engineering and Computer Science*, *13*(3), 902–909. https://doi.org/10.11591/ijeecs.v13.i3.pp902-909

Firdaus, A. A., Saputro, J. S., Anwar, M., Adriyanto, F., Maghfiroh, H., Ma'arif, A., Syuhada, F., & Hidayat, R. (2024). Application of Sentiment Analysis as an Innovative Approach to Policy Making: A Review. *Journal of Robotics and Control (JRC)*, *5*(6), 1784–1798. https://doi.org/10.18196/jrc.v5i6.22573

Godard, R., & Holtzman, S. (2022). The Multidimensional Lexicon of Emojis: A New Tool to Assess the Emotional Content of Emojis. *Frontiers in Psychology*, *13*(June), 1–10. https://doi.org/10.3389/fpsyg.2022.921388

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Hermaliani, E. H., & Ernawati, M. (2024). Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada Klasifikasi Ujaran Kebencian. *Computer Science (CO-SCIENCE)*, *4*(1), 80–88.

Idris I, Mustofa Y, & Salihi I. (2023). Analisis Sentimen Terhadap Penggunaan Aplikasi Shopee Mengunakan Algoritma Support Vector Machine (SVM). *Jambura Journal of Electrical and Electronics Engineering*, *5*, 32–35.

Indriani, F., Nugroho, R. A., Faisal, M. R., & Kartini, D. (2024). Comparative Evaluation of IndoBERT, IndoBERTweet, and mBERT for Multilabel Student Feedback Classification. *Jurnal RESTI*, *8*(6), 748–757.

https://doi.org/10.29207/resti.v8i6.6100

Kapoor, S., & Perrone, V. (2021). A Simple and Fast Baseline for Tuning Large XGBoost Models. *CoRR*, *abs/2111.06924*. https://arxiv.org/abs/2111.06924

Kim, K., & Ko, Y. (2025). Dependency Parsing with the Structuralized Prompt Template. *ArXiv Preprint ArXiv:2502.16919*.

Kosheleva, O., Gholamy, A., & Kreinovich;, V. (2018). *Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation*. https://scholarworks.utep.edu/cs_techrep/1209/

Koto, F., Lau, J. H., & Baldwin, T. (2021). IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10660–10668. https://doi.org/10.18653/v1/2021.emnlp-main.833

Kurniawan, M., Kusrini, K., & Arief, M. R. (2022). Part of Speech Tagging Pada Teks Bahasa Indonesia dengan BiLSTM + CNN + CRF dan ELMo. *Jurnal Eksplora Informatika*, *11*(1), 29–37. https://doi.org/https://doi.org/10.30864/eksplora.v11i1.506

Kusuma, A., Yudha, D. A., Afwa, M. B., & Darono, H. E. (2025). EVALUATION OF USER PERCEPTIONS AND SATISFACTION THROUGH. *Techno Nusa Mandiri: Journal of Computing and Information Technology*, *22*(2).

Lampung, P. B. P. (2025). *Gemilang 100 Hari Duet Mirza-Jihan, Kepuasan Publik Lampung Capai 83,67 Persen*. BPKAD PROVINSI LAMPUNG.

Le, T. A., Moeljadi, D., Miura, Y., & Ohkuma, T. (2016). Sentiment Analysis for Low Resource Languages: A Study on Informal Indonesian Tweets. *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, 123–131. https://aclanthology.org/W16-5415

Liu, C., Fang, F., Lin, X., Cai, T., Tan, X., Liu, J., & Lu, X. (2021). Improving sentiment analysis accuracy with emoji embedding. *Journal of Safety Science and Resilience*, *2*(4), 246–252. https://doi.org/10.1016/j.jnlssr.2021.10.003

Lou, Y., Zhou, J., Zhou, J., Ji, D., & Zhang, Q. (2024). Emoji multimodal microblog sentiment analysis based on mutual attention mechanism. *Scientific Reports*, *14*(1), 1–12. https://doi.org/10.1038/s41598-024-80167-x

Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences*, *36*(4). https://doi.org/10.1016/j.jksuci.2024.102048

Martin, M., & Nilawati, L. (2019). Recall dan Precision Pada Sistem Temu Kembali Informasi Online Public Access Catalogue (OPAC) di Perpustakaan. *Paradigma - Jurnal Komputer Dan Informatika*, *21*(1), 77–84. https://doi.org/10.31294/p.v21i1.5064

Murodi, A., & Setiyoko, Y. A. N. (2024). The Role of Social Media in Shaping Public Perception of Politics. *Formosa Journal of Applied Sciences (FJAS)*, *3*(12), 4799–4812. https://journal.formosapublisher.org/index.php/fjas

Nasrallah Taraf, S., & T. Sadiq, A. (2024). Sentiment Analysis in Arabic Text and Emoji Using Deep Learning Methods. *Iraqi Journal for Computers and Informatics*, *50*(2), 132–143. https://doi.org/10.25195/ijci.v50i2.534

Novak, P. K., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. *PLoS ONE*, *10*(12), 1–22. https://doi.org/10.1371/journal.pone.0144296

Nurhayati, L. D., & Rahardi, M. (2025). Impact of SMOTE and ADASYN on Class Imbalance in Metabolic Syndrome Classification Using Random Forest Algorithm. *Journal of Applied Informatics and Computing (JAIC)*, *9*(5), 2807–2813.

Nurwanda, N., Suarna, N., & Prihartono, W. (2024). PENERAPAN NLP (NATURAL LANGUAGE PROCESSING) DALAM ANALISIS SENTIMEN PENGGUNA TELEGRAM DI PLAYSTORE. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(2), 1841–1846. https://doi.org/10.36040/jati.v8i2.8469

Oztornaci, B., Ata, B., & Kartal, S. (2024). Analysing Household Food Consumption in Turkey Using Machine Learning Techniques. *Agris On-Line Papers in Economics and Informatics*, *16*(2), 97–105. https://doi.org/10.7160/aol.2024.160207

Permana, A. A., Fahrezi, M. F., Kristiyanti, D. A., & Sihotang, M. (2021). Sentimen Analisis Opini Masyarakat Pada Media Sosial Twitter Terhadap Vaksin Berbayar

Menggunakan Metode Naïve Bayes Classifier (Nbc). *Jurnal Teknik*, *10*(2), 84–92. https://doi.org/10.31000/jt.v10i2.5471

Proti, D., & Vuli, I. (2023). Numerical Feature Selection and Hyperbolic Tangent Feature Scaling in Machine Learning-Based Detection of Anomalies in the Computer Network Behavior. *Electronics*, *12*(9), 1–20. https://doi.org/https://doi.org/10.3390/electronics12194158

Pugliese, R., Regondi, S., & Marini, R. (2021). Machine learning-based approach: Global trends, research directions, and regulatory standpoints. *Data Science and Management*, *4*(November), 19–29. https://doi.org/10.1016/j.dsm.2021.12.002

Putra, A. H., & Salam, A. (2025). A Comparative Performance of SMOTE , ADASYN and Random Oversampling in Machine Learning Models on Prostate Cancer Dataset. *Journal of Applied Informatics and Computing (JAIC)*, *9*(3), 603–610.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *CoRR*, *abs/2003.0*. https://arxiv.org/abs/2003.07082

Rahmau'dina, R. W., Fredico, M., & Pradana, B. C. S. A. (2024). Media Sosial Instagram Sebagai Media Komunikasi Interaktif Di Grande Garden Cafe. *RELASI: Jurnal Penelitian Komunikasi (e-ISSN: 2807-6818)*, *4*(03), 11–18. https://doi.org/10.69957/relasi.v4i03.1613

Retnoningsih, E., & Pramudita, R. (2020). Mengenal Machine Learning Dengan Teknik Supervised Dan Unsupervised Learning Menggunakan Python. *BINA INSANI ICT JOURNAL*, *7*(2), 156. https://doi.org/10.51211/biict.v7i2.1422

Rivaldi, R. C., Wismarini, T. D., Lomba, J. T., & Semarang, J. (2024). Analisis Sentimen Pada Ulasan Produk Dengan Metode Natural Language Processing (NLP) (Studi Kasus Zalika Store 88 Shopee). *Jurnal Ilmiah Elektronika Dan Komputer*, *17*(1), 120–128.

Rizqilla, M., Nurani, D., Supriatin, Rahmi, A. N., & Maemunah, M. (2025). Analisis Sentimen Kinerja Jokowi Menggunakan Motode Naïve Bayes. *Remik : Riset Dan E-Jurnal Manajemen Informatika Komputer*, *9*(2), 550–561.

Sadeli, A. F., & Lawanda, I. I. (2023). Recall, Precision, and F-Measure for Evaluating Information Retrieval System in Electronic Document Management Systems

(EDMS). *Khizanah Al-Hikmah : Jurnal Ilmu Perpustakaan, Informasi, Dan Kearsipan*, *11*(2), 231–241. https://doi.org/10.24252/kah.v11i2a8

Samih, A., Ghadi, A., & Fennan, A. (2023). Enhanced sentiment analysis based on improved word embeddings and XGboost. *International Journal of Electrical and Computer Engineering*, *13*(2), 1827–1836. https://doi.org/10.11591/ijece.v13i2.pp1827-1836

Saputri, M. S., Mahendra, R., & Adriani, M. (2018). Emotion Classification on Indonesian Twitter Dataset. *2018 International Conference on Asian Language Processing (IALP)*, 90–95. https://doi.org/10.1109/IALP.2018.8629262.

Sathyanarayanan, S. (2024). Confusion Matrix-Based Performance Evaluation Metrics. *African Journal of Biomedical Research*, *December*, 4023–4031. https://doi.org/10.53555/ajbr.v27i4s.4345

Sihombing, T. S. R., Asmara, S., & Kurniawati, D. (2024). The Influence of Public Communication Through Instagram on Plastic Waste Management on Community Knowledge and Behavior in Medan Tuntungan Sub-District, Medan City. *Indonesian Journal of Advanced Research*, *3*(1), 143–156. https://doi.org/10.55927/ijar.v3i1.7917

Sinaga, H. H., & Agustian, S. (2022). Pebandingan Metode Decision Tree dan XGBoost untuk Klasifikasi Sentimen Vaksin Covid-19 di Twitter. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *8*(3), 107–114. https://doi.org/10.25077/teknosi.v8i3.2022.107-114

Surendro, K., Setiawan, Y., & Maulidevi, N. U. (2024). The Optimization of n-Gram Feature Extraction Based on Term Occurrence for Cyberbullying Classification. *Data Science Journal*, *23*(31), 1–21. https://doi.org/10.5334/dsj-2024-031

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *Applied Sciences (Switzerland)*, *13*(7). https://doi.org/10.3390/app13074550

Tarwidi, D., Redjeki, S., Adytia, D., & Apri, M. (2023). An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach. *MethodsX*, *10*(December 2022). https://doi.org/10.1016/j.mex.2023.102119

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł.,

& Polosukhin, I. (2017). Attention Is All You Need. *International Conference on Information and Knowledge Management, Proceedings*, *Nips*, 4752–4758. https://doi.org/10.1145/3583780.3615497

Xu, Q. A., Chang, V., & Jayne, C. (2022). A systematic review of social media-based sentiment analysis: Emerging trends and challenges. *Decision Analytics Journal*, *3*(May), 100073. https://doi.org/10.1016/j.dajour.2022.100073

Yulistiani, S. (2024). Analisis Sentimen Terhadap Calon Presiden Indonesia 2024 dengan Metode Extreme Gradient Boosting (XGBOOST). *Jurnal Informatika: Jurnal Pengembangan IT*, *9*(3), 322–328. https://doi.org/10.30591/jpit.v9i3.6127