SKRIPSI

ANALISIS SENTIMEN KOMENTAR BERBASIS TEKS DAN EMOJI PADA VIDEO KONTEN EDUKASI YOUTUBE MENGGUNAKAN METODE BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT)

Disusun dan diajukan oleh:

ILHAM
D121 17 1309

PROGRAM STUDI SARJANA TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS HASANUDDIN
GOWA
2024

LEMBAR PENGESAHAN SKRIPSI

ANALISIS SENTIMEN KOMENTAR BERBASIS TEKS DAN EMOJI PADA VIDEO KONTEN EDUKASI YOUTUBE MENGGUNAKAN METODE BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT)

Disusun dan diajukan oleh

ILHAM
D121171309

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka
Penyelesaian Studi Program Sarjana Program Studi
Fakultas Teknik Universitas Hasanuddin
Pada tanggal 31 Juli 2024
dan dinyatakan telah memenuhi syarat kelulusan

Menyetujui,

Pembimbing Utama,	Pembimbing Pendamping,
Dr.Eng. Zulkifli Tahir, S.T., M.Sc.	Ir. Christoforus Yohannes, M.T.
NIP 198404032010121004	NIP 196007161987021002

Ketua Program Studi,

Prof. Dr. Ir. Indrabayu, ST., MT., M.Bus.Sys., IPM, ASEAN. Eng.
NIP 197507162002121004

PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini ;
Nama : ILHAM
NIM : D121 17 1309
Program Studi : Teknik Informatika
Jenjang : S1

Menyatakan dengan ini bahwa karya tulisan saya berjudul

Analisis Sentimen komentar berbasis teks dan emoji pada video konten edukasi youtube Menggunakan Metode Bidirectional Encoder Representations from Transformers (BERT)

Adalah karya tulisan saya sendiri dan bukan merupakan pengambilan alihan tulisan orang lain dan bahwa skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri.

Semua informasi yang ditulis dalam skripsi yang berasal dari penulis lain telah diberi penghargaan, yakni dengan mengutip sumber dan tahun penerbitannya. Oleh karena itu semua tulisan dalam skripsi ini sepenuhnya menjadi tanggung jawab penulis. Apabila ada pihak manapun yang merasa ada kesamaan judul dan atau hasil temuan dalam skripsi ini, maka penulis siap untuk diklarifikasi dan mempertanggungjawabkan segala resiko.

Segala data dan informasi yang diperoleh selama proses pembuatan skripsi, yang akan dipublikasi oleh Penulis di masa depan harus mendapat persetujuan dari Dosen Pembimbing.

Apabila dikemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan isi skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Gowa, 31 JULI 2024

Yang Menyatakan

ILHAM

ABSTRAK

ILHAM. *Analisis Sentimen komentar berbasis teks dan emoji pada video konten edukasi Youtube Menggunakan Metode Bidirectional Encoder Representations from Transformers (BERT)* (dibimbing oleh Zulkifli Tahir dan Christoforus Yohannes)

Pada zaman digital saat ini, *Youtube* telah menjadi platform video paling populer di dunia. Salah satu konten yang mengalami pertumbuhan yaitu konten edukasi dimana konten ini sangat penting untuk diperhatikan agar informasi yang diberikan bisa dipelajari dengan baik dan tidak menyesatkan.

Beragam reaksi pengguna Ketika menonton video sangat memengaruhi reputasi konten tersebut. Dalam menilai video, *Youtube* memiliki fitur komentar, namun seringkali komentar yang mencapai ribuan sangat sulit jika di analisa secara manual, sehingga dibutuhkan pendekatan analisis sentimen untuk dapat mengetahui Kesimpulan dari video secara efisien.

Metodologi yang digunakan pada penelitian ini meliputi pengumpulan dataset komentar video, kemudian menerapkan fine tunning model IndoBERT untuk tugas analisis sentimen. Pengujian dan evaluasi model dilakukan dalam 3 skenario berbeda dengan ukuran batch 8, 16, dan 32, kemudian mengukur performa berdasarkan precision, recall, dan f1-score.

Penelitian menunjukkan bahwa model yang dikembangkan mampu melakukan sentimen analisis dengan performa yang baik, skor tertinggi yaitu pada scenario pertama dengan *batch size* 8 *epoch* ke 4. dengan nilai akurasi sebesar 85%

Kata Kunci: Analisis Sentimen, NLP, BERT, indoBERT, Video konten edukasi

ABSTRACT

ILHAM. *Analisis Sentimen komentar berbasis teks dan emoji pada video konten edukasi Youtube Menggunakan Metode Bidirectional Encoder Representations from Transformers (BERT)* (dibimbing oleh Zulkifli Tahir dan Christoforus Yohannes)

In the current digital era, *Youtube* has become the most popular video platform in the world. One type of content experiencing growth is educational content, which is very important to monitor to ensure the information provided can be learned effectively and is not misleading.

The diverse reactions of users when watching videos greatly affect the reputation of the content. In evaluating videos, *Youtube* has a comment feature, but often comments numbering in the thousands are very difficult to analyze manually, thus requiring a sentiment analysis approach to efficiently understand the conclusions from the videos.

The methodology used in this study includes collecting a dataset of video comments, then applying fine-tuning of the IndoBERT model for sentiment analysis tasks. Testing and evaluation of the model were conducted in three different scenarios with batch sizes of 8, 16, and 32, and performance was measured based on precision, recall, and F1-score.

The research shows that the developed model is capable of performing sentiment analysis with good performance, with the highest score in the first scenario with a batch size of 8 at the 4th epoch, achieving an accuracy of 85%.

Keywords: Sentiment Analysis, NLP, BERT, IndoBERT, Educational Video Content

DAFTAR ISI

LEMBAR PENGESAHAN SKRIPSI i
PERNYATAAN KEASLIAN ii
ABSTRAK iii
ABSTRACT iv
DAFTAR ISI v
DAFTAR GAMBAR vii
DAFTAR TABEL viii
DAFTAR SINGKATAN DAN ARTI SIMBOL ix
KATA PENGANTAR x
BAB 1 PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusah Masalah 3
1.3 Tujuan Penelitian 3
1.4 Manfaat Penelitian 3
1.5 Ruang Lingkup 3
BAB 2 TINJAUAN PUSTAKA 5
2.1 Youtube 5
2.2 *Natural Language Processing* (NLP) 6
2.3 Analisis Sentimen 7
2.4 Text Mining 8
2.5 Transformer 8
2.6 BERT 9
2.7 IndoBERT 15
2.8 Multiclass Confusion Matrix 15
2.9 Batch Size 17
2.10 Training & Validation Loss 17
2.11 FastAPI 18
BAB 3 METODE PENELITIAN 20
3.1. Waktu dan Lokasi Penelitian 20
3.2. Instrument Penelitian 20
3.3. Tahapan Penelitian 21
3.4. Perancangan Sistem 23

3.5. Scrapping Data (Input Dataset) 24
3.6. Preprocessing Data 24
3.7. Pelabelan data 26
3.8. Splitting Data 27
3.9. Pelatihan Menggunakan Model Pre-trained BERT 27
3.10. Evaluasi Model 27
BAB 4 HASIL DAN PEMBAHASAN 29
4.1. Pelatihan dan validasi Model 29
4.2. Perbandingan kinerja model 38
4.3. Pengujian Model 41
4.4. Penerapan Model menggunakan framework FastAPI 42
5.1 Kesimpulan 46
5.2 Saran 46
DAFTAR PUSTAKA 47

DAFTAR GAMBAR

Gambar 1 Analisis Sentimen menggunakan Machine Learning 8
Gambar 2 Analisis Sentimen menggunakan Deep Learning 8
Gambar 3Arsitektur Transformer (Vaswani, 2017) 9
Gambar 4 Perbandingan BERT-Base dan BERT- Large 10
Gambar 5. Arsitektur Encoder pada Transformer (Vaswani, 2017) 10
Gambar 6 Input Embedding Layer (Kenton et al., 2018) 11
Gambar 7 Arsitektur *Pra-trained* BERT (Kenton et al., 2018).....12
Gambar 8 Arsitektur fine tune BERT (Kenton et al., 2018).....13
Gambar 9. Classification layer pada BERT 14
Gambar 10 Multiclass Confusion Matrix 15
Gambar 11 Tahapan penelitian 21
Gambar 12 Rancangan Sistem 23
Gambar 13 Data hasil crawling 24
Gambar 14 Training Loss & Validation Loss Skenario Pertama 29
Gambar 15 F1 Score Skenario Pertama.....30
Gambar 16 Confusion Matrix Model Pelatihan Skenario Pertama pada epoch ke-4 31
Gambar 17 Training Loss & Validation Loss Skenario Kedua 33
Gambar 18 F1 Score Skenario Kedua 33
Gambar 19 Confusion Matrix Model Pelatihan Skenario Kedua pada epoch ke-4.....34
Gambar 20 Training Loss & Validation Loss Skenario Ketiga 36
Gambar 21 F1 Score Skenario Ketiga 36
Gambar 22 Confusion Matrix Model Pelatihan Skenario Ketiga pada epoch ke-3 37
Gambar 23 Grafik perbandingan Training Loss 39
Gambar 24 Grafik Perbandingan Validation Loss 40
Gambar 25 Perbandingan F1-Score tiap scenario.....40
Gambar 26 FastAPI input youtube URL 43
Gambar 27 Hasil download youtube comment.....43
Gambar 28 Hasil analisis sentimen model 44
Gambar 29 FastAPI output score comment youtube sentiment 44
Gambar 30 Tampilan website analisis sentiment BERT 44
Gambar 31 Tampilan score website analisis sentiment BERT.....45

DAFTAR TABEL

Tabel 1 contoh hasil labeling data 26
Tabel 2 Skenario Pengujian.....28
Tabel 3 Evaluasi model skenario pertama 32
Tabel 4 Evaluasi model skenario kedua 35
Tabel 5 Evaluasi model skenario ketiga 38
Tabel 6 Contoh hasil sentiment analisis 41

DAFTAR SINGKATAN DAN ARTI SIMBOL

Lambang/Singkatan	Arti dan Keterangan
BERT	*Bidirectional Encoder Representative from Transformers*
TP	*True Positive*
FP	*False Negative*
TN	*True Negative*
FN	*False Negative*
MLM	*Masked Language Model*
WER	*Word Error Rate*
NSP	*Next Sentence Prediction*

KATA PENGANTAR

Assalamualaikum Warahmatullahi Wabarakatuh

Segala puji dan syukur penulis penjatkan kehadirat Allah SWT yang telah melimpahkan karunia, hidayah, kekuatan dan pencerahan kepada penulis sehingga penulis dapat menyelesaikan tugas akhir dengan judul "Analisis Sentimen komentar berbasis teks dan emoji pada video konten edukasi *Youtube* Menggunakan Metode Bidirectional Encoder Representations from Transformers (BERT)". Sebagai syarat dalam menyelesaikan studi jenjang Strata-1 di Departemen Teknik Informatika Fakultas Teknik, Universitas Hasanuddin.

Penulis sadar benar bahwa dalam penyusunan dan penulisan laporan penelitian ini masih banyak kekurangan dan jauh dari kesempurnaan. Namun berkat dukungan, bantuan, bimbingan, kerja sama dari berbagai pihak dan Allah SWT sehingga kendala-kendala saat proses pengerjaan skripsi ini dapat diatasi dengan usaha dan kemampuan yang penulis miliki.

Selesainya skripsi ini tentunya tidak lepas dari bimbingan serta bantuan dari berbagai pihak. Penulis mengungkapkan terima kasih dan penghargaan setinggi-tingginya kepada semua pihak yang telah membantu, dan dalam kesempatan ini penulis ingin menyampaikan banyak terima kasih yang sebesar-besarnya kepada :

1. Allah S.W.T karena atas semua berkat, karunia serta pertolongan-Nya yang tiada batas, yang telah diberikan kepada Penulis disetiap langkah dalam pembuatan tugas akhir ini hingga penulisan tugas akhir ini;
2. Kedua orang tua Penulis, Bapak Nurkhamsi dan Ibu Sumarni yang selalu memberikan dukungan, doa dan semangat;
3. Dr. Eng. Zulkifli Tahir, ST., M.SC. selaku pembimbing I dan Bapak Ir. Christoforus Yohannes, M.T., selaku pembimbing II, yang senantiasa menyediakan waktu, tenaga, pikiran dan perhatian yang luar biasa untuk mengarahkan Penulis dalam penyusunan tugas akhir ini;
4. Bapak Prof. Dr. Ir. Indrabayu, S.T., M.T., M.Bus.Sys., IPM, ASEAN. Eng., selaku Ketua Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin atas bimbingannya selama masa perkuliahan Penulis;

5. Muh. Ikbal yang membantu Penulis dalam memberikan dukungan dan pemikiran-pemikiran dalam penyelesaian tugas akhir;
6. Bapak Robert, Bapak Zainuddin dan Ibu arizha serta segenap staff Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin yang telah membantu kelancaran penyelesaian tugas akhir Penulis;
7. Keluarga besar RECOGN17ER yang menjadi rekan seperjuangan selama berada di kampus sebagai anak teknik;
8. Seluruh pihak yang tidak sempat disebutkan satu persatu yang telah banyak meluangkan tenaga, waktu dan pikiran selama penyusunan laporan tugas akhir ini.

Akhir kata, penulis berharap semoga Allah SWT. Berkenan membalas segala kebaikan dari semua pihak yang telah banyak membantu. Semoga Tugas Akhir ini dapat memberikan manfaat bagi pengembangan ilmu, Aamiin.

Wassalam

Makassar, 31 Juli 2024

Penulis.
Ilham

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Pada zaman digital saat ini, seseorang dapat dengan mudahnya mengakses internet untuk menggunakan platform media sosial. Salah satu platform yang paling sering digunakan adalah *Youtube*. *Youtube* menjadi platform video paling populer di dunia. Berdasarkan data di Situs *Dataindonesia.id*, pengguna aktif *Youtube* di dunia mencapai 2,41 miliar pada kuartal II/2022, Adapun pengguna *Youtube* di indonesia menempati posisi ketiga di dunia. Berdasarkan laporan *We Are Social*, jumlahnya mencapai 127 juta pengguna (Shilvina Widi, 2022). Meningkatnya pengguna berbanding lurus dengan bertambah banyak nya konten yang di unggah di *Youtube*.

Youtube menawarkan berbagai jenis konten menarik seperti konten komedi, edukasi, musik, vlog, dll. Salah satu kategori yang mengalami pertumbuhan adalah konten video edukasi. Konten edukasi merupakan sebuah konten yang ditujukan untuk menyampaikan informasi yang khususnya berupa materi yang diyakini dapat menambah ilmu bagi audiensnya (Andiana Moedasir, 2022). Konten edukasi mencakup banyak hal. Biasanya berisi tentang materi pembelajaran, kesehatan, tutorial, riview produk, dll. Melalui media pembelajaran menggunakan *Youtube*, penonton dapat memahami materi lebih cepat daripada mempelajari melalui buku. Konten Edukasi juga menjadi hal yang paling penting untuk diperhatikan karena data dan ilmu pengetahuan yang disampaikan kepada pengguna harus dapat dipelajari dengan baik dan tidak menyesatkan.

Beragam reaksi pengguna ketika melihat konten video yang di unggah di *Youtube* sangat mempengaruhi reputasi konten video dan channel tersebut. Dalam menilai konten video, *Youtube* memiliki 2 fitur, yaitu fitur *like* dan *dislike* memungkinkan pengguna untuk mengevaluasi konten video dalam angka, namun pada tanggal 10 November 2021, pengguna tidak dapat lagi melihat perbandingan rasio *like* dan *dislike* dikarenakan *Youtube* secara resmi mengumumkan bahwa pengguna tidak dapat lagi melihat jumlah *dislike* yang terdapat pada video. Fitur lain yang dimiliki *Youtube* yang memungkinkan pengguna mengetahui tanggapan pengguna lain adalah komentar, komentar berupa teks mampu merepresentasikan

pendapat pengguna lain ketika melihat konten video. Pengguna dapat melihat tanggapan pengguna lain dengan cara menganalisa komentar pada sebuah video, namun video populer biasanya memiliki komentar yang mencapai ribuan yang membuat menganalisa secara manual menjadi sulit untuk dilakukan. Untuk dapat memahami persepsi dari sebuah komentar terhadap suatu konten video dilakukan pendekatan berupa analisis sentimen. Analisis sentimen merupakan ilmu yang berguna untuk menganalisis pendapat seseorang, sentiment seseorang, evaluasi seseorang, sikap seseorang dan emosi seseorang ke dalam bahasa tertulis. Komentar yang diberikan oleh pengguna terhadap sebuah konten video dapat dikategorikan ke dalam sebuah jenis sentimen. Pada umumnya, jenis sentimen dibagi menjadi tiga jenis yaitu sentimen positif, negatif, dan netral. Namun tidak menutup kemungkinan terdapat beberapa jenis lain dalam pengelompokan sentimen, misalnya sentimen berbasis emosi seperti marah, sedih, kecewa, dan sejenisnya. Saat ini, terdapat banyak pengguna yang mengeskspresikan pesan yang disampaikan dengan menggunakan emoji atau sebagai pendukung dari komentar yang dituliskan. Penggunaan emoji juga akan mempengaruhi hasil sentimen dari sebuah komentar. Dengan menganalisa emoji pada teks, akan lebih mudah memahami maksud dari komentar yang disampaikan pengguna.

Penelitian ini dilakukan untuk menganalisa sentimen pada konten video edukasi *Youtube* berdasarkan komentar teks dan emoji. Pada penelitian ini, analisis sentimen dilakukan menggunakan metode IndoBERT, IndoBERT merupakan model versi Indonesia dari *Bidirectional Encoder Representations from Transformers (BERT)* dikarenakan BERT dapat memahami konteks dari komentar secara mendalam tentang semantik kata yang ada. BERT merupakan teknik *machine learning* yang berbasis *transformer* untuk *natural language processing* (NLP) yang dikembangkan oleh Google. dirancang agar dapat melakukan *pre-train* pada teks tanpa label secara dua arah dengan menggunakan teks disekitar kata yang memungkinkan BERT untuk mengetahui konteks dari sebuah kalimat. (Devlin. Dkk, 2019).

Oleh karena itu, penulis mengangkat penelitian dengan judul "Analisis Sentimen Komentar Berbasis Teks dan Emoji pada Video Konten Edukasi di *Youtube* Menggunakan Metode *Bidirectional Encoder Representations From*

Transformers (BERT)". Penelitian ini dimaksudkan untuk menganalisis sentimen pada komentar sebuah video agar dapat mengklasifikasi video berdasarkan opini pengguna dengan menggunakan metode BERT.

1.2 Rumusah Masalah

Berdasarkan latar belakang masalah yang telah dijelaskan di atas maka rumusan masalah dalam penelitian ini adalah:

1. Bagaimana performa model analisis sentimen terhadap komentar teks dan emoji menggunakan BERT
2. Bagaimana merancang website analisis sentimen komentar *Youtube* menggunakan model BERT

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Menganalisis akurasi sentimen terhadap komentar teks dan emoji menggunakan BERT
2. Menghasilkan website analisis sentimen komentar *Youtube* menggunakan model BERT

1.4 Manfaat Penelitian

Dengan dilakukannya penelitian ini, diharapkan manfaat yang didapatkan antara lain:

1. Mempermudah dalam pengambilan Keputusan untuk memilih konten edukasi di *Youtube* yang dapat digunakan oleh pelajar.
2. Hasil penelitian dapat dijadikan sebagai rujukan penelitian terkait.

1.5 Ruang Lingkup

1. Kumpulan data bersumber dari media sosial *Youtube*
2. Data yang digunakan berupa komentar berbahasa Indonesia pada video konten edukasi yang berasal dari *Youtube*.
3. Komentar video yang di ambil maksimal 3000 komentar per video

4. Video yang dianalisis merupakan video konen edukasi yang secara khusus memberikan panduan tata cara dan informasi.
5. Klasifikasi terdiri atas tiga kelas, yaitu positif, negatif dan netral.
6. Metrik Evaluasi pengukuran performa menggunakan *Confusion Matrix* yang terdiri dari :
   - *Precision*
   - *Recall*
   - *F1-Score*

BAB 2
TINJAUAN PUSTAKA

2.1 Youtube

Youtube merupakan situs jejaring media sosial yang berisikan berbagai jenis video, pengguna *Youtube* bisa menonton dan mengupload video lewat situs tersebut, banyak sekali jenis video yang terdapat dari *Youtube* mulai dari hiburan, wawasan, serta berita, (Wiryany 2019), *Youtube* merupakan situs video sharing yang dimiliki google inc. dikategorikan sebagai media yang berisikan jutaan video. *Youtube* dibangun pada Tahun 2005 merupakan titik awal dari lahirnya situs video upload *Youtube.com* yang didirikan oleh 3 orang karyawan perusahaan finance online *PayPal* di Amerika Serikat, mereka adalah Chad Hurley, Steve Chen, dan Jawed Karim.

Youtube sendiri terinspirasi dari nama Sebuah kedai pizza dan restoran Jepang di San Mateo, California. Yang kemudian dibeli sahamnya oleh perusahaan google hingga berkembang seperti sekarang, *Youtube* memiliki potensi yang cukup baik dijadikan sebuah media pembelajaran karena dijaman sekarang banyak sekali genre video yang terdapat di youtub terutama video yang berisikan konten pendidikan, contohnya video sejarah berbentuk dokumenter ataupun animasi.

Ada beberapa kelebihan dan kekurangan *Youtube* (Wiryany, 2019) kelebihan dari *Youtube* diantaranya:

1) Memudahkan pengguna dalam mecari suatu topik yang ingin dicari melalui kata kunci untuk memunculkan banyak hal yang berkaitan dengan topik tersebut berupa video yang telah tersedia di dalamnya
2) Konten yang ada di *Youtube* lebih beragam dibandingkan dengan media masa lainya seperti *Instagram, Twitter, Facebook*, karena video memiliki ruang pemutaran yang lebih besar
3) *Youtube* bisa dijadikan sebagai sarana belajar otodidak, dengan adanya *Youtube* memudahkan seseorang untuk menggali suatu hal yang belum di pahami

Walaupun *Youtube* memiliki berbagai kelebihan *Youtube* juga memiliki kekurangannya diataranya:

1) Kebebasan dalam mengakses konten yang cukup berbahaya bagi anak anak dibawah umur atau konten yang yang kurang cocok bagi beberapa kalangan orang
2) Beberapa konten yang bermuatan berita berpotensi hoak atau unsur berita pembohong yang memicu konflik
3) Tergesernya televisi karena beragamnya konten *Youtube* yang dimiliki membuat orang-orang lebih tertarik.

2.2 *Natural Language Processing* (NLP)

Natural Language Processing atau Pengolahan Bahasa Alami adalah salah satu cabang ilmu Kecerdasan Buatan yang mempelajari dan mengembangkan bagaimana komputer dapat mengerti, memahami, dan memproses bahasa alami dalam bentuk teks atau tuturan kata. NLP menganalisa bahasa manusia sedemikian rupa sehingga komputer dapat memahami bahasa alami seperti halnya manusia (Ghosh et al., 2012). NLP adalah salah satu bidang antar disiplin yang menggabungkan komputasi linguistik, ilmu komputasi, ilmu kognitif, dan kecerdasan buatan. Pada umumnya, NLP banyak diaplikasikan di berbagai hal seperti *speech recognition*, pemahaman bahasa lisan, sistem dialog, analisis leksikal, mesin penerjemah, *knowledge graph*, analisis sentimen, sistem pintar dan peringkasan bahasa alami.

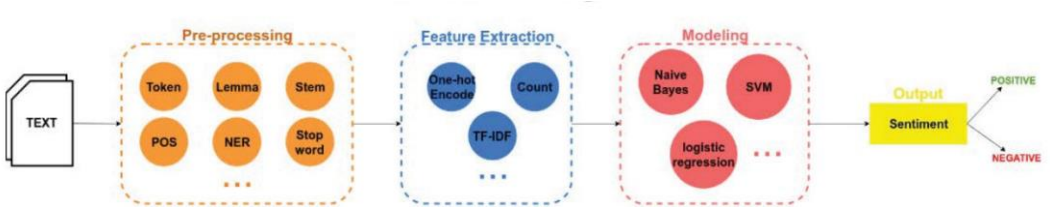
Sebuah sistem NLP dapat dimulai dari tingkat kata untuk menentukan struktur dan sifat morfologis (seperti *part-of-speech* atau makna) dari kata; kemudian dapat beralih ke tingkat kalimat untuk menentukan urutan kata, tata bahasa, dan arti dari seluruh kalimat. Kemudian ke konteks dan keseluruhan domain. Kata atau kalimat yang diberikan mungkin memiliki makna atau konotasi yang berbeda dalam konteks tertentu, yang terkait dengan banyak kata atau kalimat lain dalam konteks yang diberikan.

2.3 Analisis Sentimen

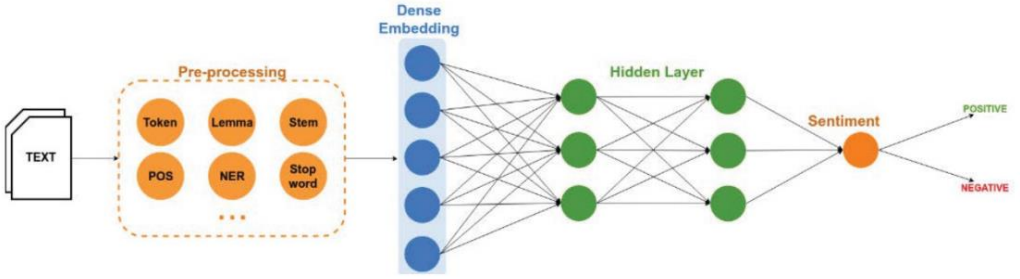
Analisis sentimen adalah suatu bidang penelitian yang terus berkembang dan berhubungan dengan berbagai disiplin ilmu seperti Data Mining, *Natural Language Processing* (NLP), dan *Machine Learning*. Fokus dari analisis sentimen adalahuntuk mengekstraksi sentimen dari sebuah kalimat berdasarkan kontennya. *Sentiment analysi*s, juga dikenal sebagai *opinion mining*, merupakan proses yang dilakukan secara otomatis untuk memahami, mengekstrak, dan memproses data teks guna mendapatkan informasi yang terkandung dalam suatu kalimat opini. Tujuan dari analisis sentimen ini adalah untuk melihat pendapat atau kecenderungan opini seseorang terhadap suatu masalah atau objek, apakah memiliki kecenderungan positif, negatif, atau netral (Halim & Safuwan, 2023).

Metode *supervised machine learning* seperti *Naïve Bayes* (NB), *Support Vector Machine* (SVM), atau *Logistic Regression* (LR) telah digunakan sebagai upaya untuk menyelesaikan masalah pengkategorian sentimen teks. Namun, algoritma *machine learning* klasik ini menunjukkan kinerja yang kurang memuaskan saat berhadapan dengan *cross-lingual* atau *cross-domain* data. Berbeda dengan pendekatan berbasis *deep learning* yang menunjukkan kinerja yang lebih unggul dibandingkan dengan algoritma klasik tersebut.

Dengan menggabungkan struktur *multi-layer* ke dalam *hidden layers* pada *neural network's*, *deep learning* dapat mencapai hasil yang lebih kompleks. Metode konvensional *machine-learning* memerlukan fitur yang ditentukan dan diambil secara manual atau melalui teknik seleksi fitur. Sebaliknya, model *deep learning* secara otomatis mempelajari dan mengekstrak informasi, meningkatkan akurasi dan kinerja. Perbandingan antara standar *machine learning* seperti SVM, *Naïve Bayes*, dan *decision tree* dengan *deep learning* untuk analisis sentimen dapat dilihat pada Gambar 1 dan Gambar 2. Untuk menyelesaikan masalah yang sulit seperti di bidang gambar, pengenalan suara, dan NLP, solusi terbaik adalah menggunakan *deep learning* (Kalluri, 2023).



Gambar 1 Analisis Sentimen menggunakan Machine Learning



Gambar 2 Analisis Sentimen menggunakan Deep Learning

2.4 Text Mining

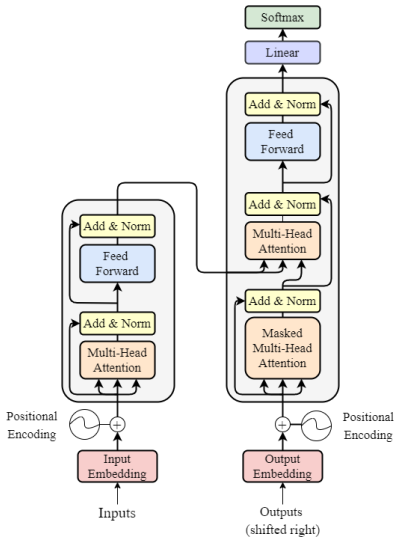
Text mining merupakan suatu proses yang menggunakan statistik, matematika, kecerdasan buatan, dan pembelajaran mesin untuk mengambil dan mengidentifikasi informasi yang bernilai. Definisi dari data mining adalah proses penemuan pola dalam data. Berdasarkan tujuannya, data mining dapat dikelompokkan menjadi deskripsi, estimasi, prediksi, klasifikasi, pengelompokan, dan asosiasi (Ikhromr et al., 2023). Data mining dapat juga diartikan sebagai suatu proses yang menggunakan satu atau lebih teknik pembelajaran mesin (*machine learning*) untuk menganalisis dan menghasilkan pengetahuan secara otomatis (Fauzan et al., 2023). Kehadiran data mining menjadi penting karena ada masalah ledakan data yang terjadi belakangan ini. Banyak organisasi telah mengumpulkan jumlah data yang sangat besar selama bertahun-tahun (Susanto & Sudiyatno, 2014).

2.5 Transformer

Transformer merupakan arsitektur *neural network* yang sepenuhnya berlandaskan pada *attention mechanism*. Berbeda dengan model tradisional yang memproses data secara berurutan, Transformer dapat memahami seluruh teks sekaligus, memungkinkan untuk memahami konteks dengan lebih efisien. Hal ini

membuatnya sangat efektif untuk tugas-tugas seperti penerjemahan mesin dan analisis sentimen.

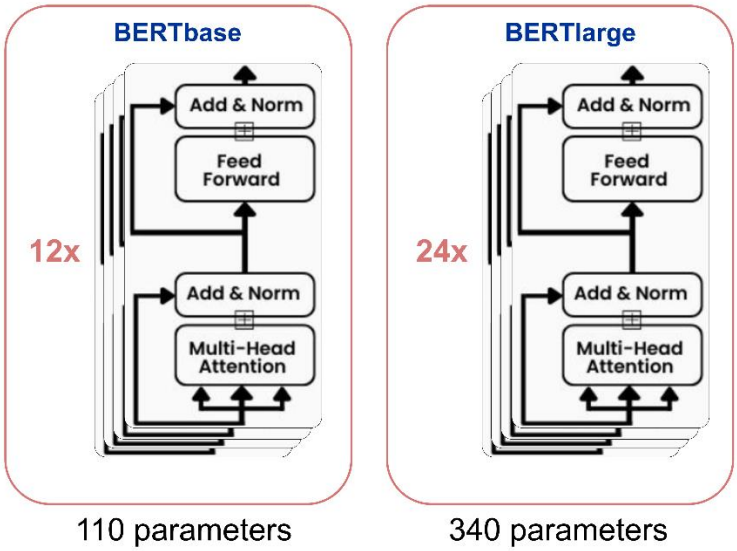
Dalam penelitian ini, model sentiment analisis dibuat menggunakan model yang dibangun berdasarkan arsitektur *Transformer*. Model *Transfomer* dapat dilihat pada gambar 3 terdiri dari dua bagian yaitu *encoder* yang berada disebelah kiri dan *decoder* yang berada di sebelah kanan (Vaswani, 2017).



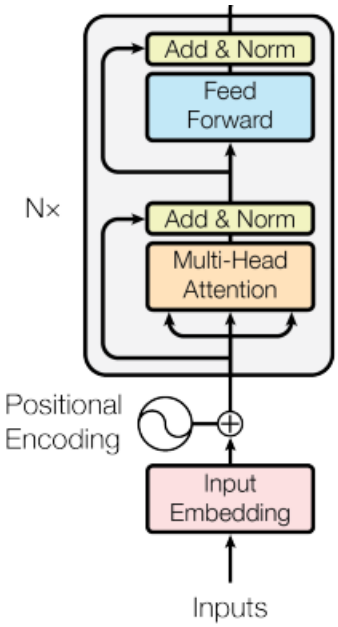
Gambar 3Arsitektur Transformer (Vaswani, 2017)

2.6 BERT

BERT dikembangkan pada tahun 2018 untuk membantu tugas komputer memahami bahasa alami dalam memahami konteks frasa (Kenton et al., 2019). BERT sesuai dengan namanya *Bidirectional Encoder Representations from Transformers* merupakan resresentasi *encoder* dari *Transformer*, tujuan dari BERT adalah untuk membuat sebuah *Language Model* oleh karena itu BERT hanya membutuhkan bagian *encoder* dari *Transformer* dan tidak memerlukan bagian *decoder*. Model BERT memiliki dua versi yaitu BERT-Base dan BERT-Large yang memiliki perbedaan pada ukuran model, dapat dilihat pada gambar 4 BERT-Base memiliki 12 *layer* atau *encoder block* yang disusun dan memiliki 110 juta *parameters* dan BERT-Large memiliki 24 *layer* atau *encoder block* yang disusun dan memiliki 340 juta *parameters*. Pada penilitian ini versi model BERT yang digunakan adalah versi BERT-Base dengan 12 *layers*.

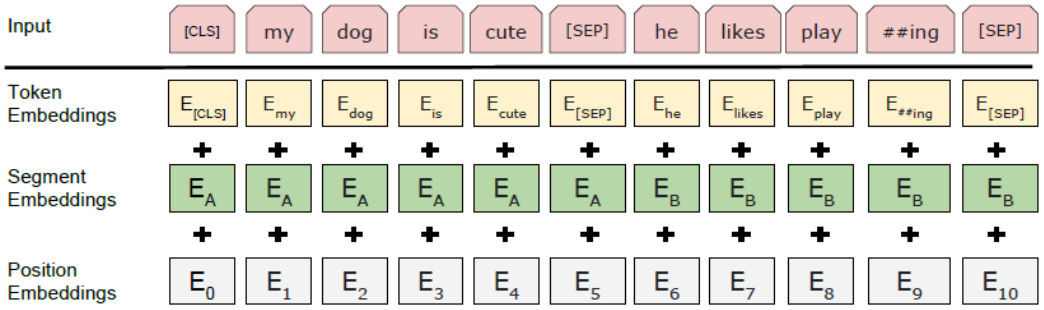


Gambar 4 Perbandingan BERT-Base dan BERT- Large



Gambar 5. Arsitektur Encoder pada Transformer (Vaswani, 2017)

Setiap *layers* atau *encoder* pada model BERT terdiri dari 2 bagian utama yaitu *multi-head attention*, dan *feed forward*. Dan sebelum data *input* masuk dan diproses di setiap *layer* data *input* akan melelaui proses *input embedding* yang dapat dilihat pada gambar 6 *Input embedding* pada BERT merupakan kombinasi dari tiga jenis *embedding* yang berbeda yaitu *token embeddings*, *segment embeddings*, dan *positional embeddings*.



Gambar 6 Input Embedding Layer *(Kenton et al., 2018)*

Dapat dilihat pada gambar 6 Pada *token embeddings* input token diubah menjadi vektor yang menyimpan data semantik, *token embedding* pada *BERT* menggunakan *pre-trained embeddings* yaitu *WordPieces* yang memiliki 30K *vocabulary tokens*, kemudian pada *segment embedding* input token diberi kode untuk membedakan kalimat satu dengan kalimat lainnya berdasarkan token [SEP] dan dapat dilihat pada gambar 6 token dengan kode A dan token dengan kode B dipisahkan oleh token [SEP], dan yang terakhir pada *positional embeddings* token diberi kode untuk mengetahui posisi token pada input data. Dan dengan menggabungkan tiga jenis *embeddings* akan menghasilkan *input embedding* yang akan masuk ke *layer* atau *encoder* pada model *BERT*.

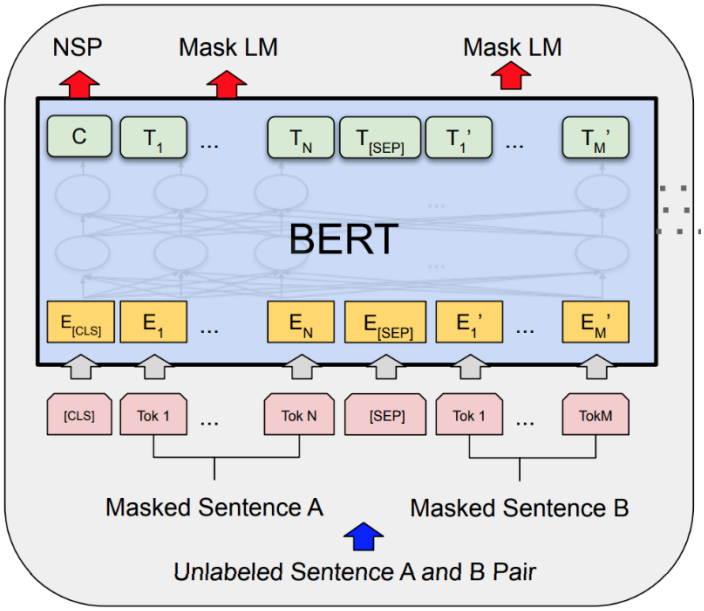
Pada *layer encoder*, *input embedding* pertama melalui mekanisme *multi-head attention* dimana setiap *token* dibuatkan representasi *attention vectors* yang memiliki informasi relasi kontekstual setiap kata pada kata yang lain, setelah melewati *multi-head attention* dilakukan proses *Add & Norm* dimana proses ini merujuk pada dua langkah proses yang dikenal sebagai *residual connection* dan *layer normalization*, pada *residual connection* output dari *sub-layer* dalam hal ini adalah output dari *multi-head attention* ditambahkan kembali dengan *input* asli dengan tujuan untuk menghindari masalah yang disebut *vanishing gradient*, setelah penambahan tersebut normalisasi dilakukan dengan menyesuaikan skala dan translasi dari output agar memiliki *mean* 0 dan *varians* 1, hal ini untuk membantu memastikan bahwa dari setiap *sub-layer* memiliki rentang nilai yang konsisten, yang membantu model belajar dengan lebih stabil dan efektif.

Setelah proses *multi-head attention* output yang berupa *attention vectors* melalui *sub-layer feed forward* dimana pada proses ini *attention vectors* diubah menjadi bentuk yang lebih mudah untuk di proses pada *layer* selanjutnya. Pada

model BERT-Base yang memiliki 12 *layer* proses ini akan dilakukan sebanyak 12 kali sesuai dengan ukuran dari model.

Pada penelitian ini BERT digunakan untuk menyelesaikan tugas analisis sentimen, dan untuk menyelesaikan tugas tersebut terdapat dua tahap yaitu tahap pertama BERT perlu dilatih agar dapat memahami bahasa dengan baik kemudian tahap kedua dilakukan *fine tune* untuk mengajarkan BERT tugas analisis sentimen.

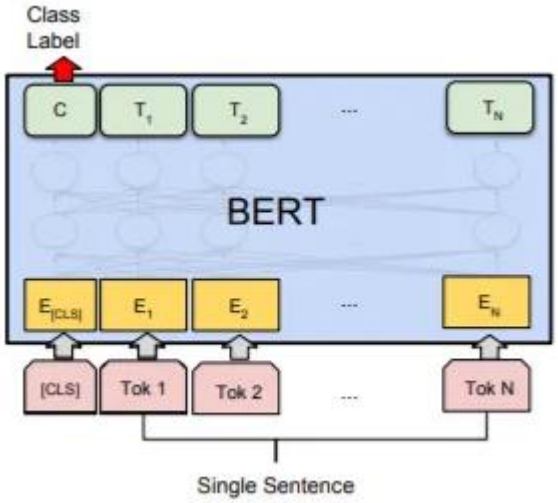
Pada tahap pertama yaitu tahap *pre-trained* BERT kita memerlukan agar model BERT dapat memahami bahasa dengan baik, untuk mencapai tujuan tersebut model BERT dilatih dengan menggunakan *unlabeled datasets* yang besar untuk menyelesaikan dua tugas yaitu *Masked Language Model* (MLM) dan *Next Sentence Prediction* (NSP). Dimana pada tugas MLM *input* token secara acak diberi *mask* untuk menyembunyikan nilai asli dari *token* tersebut dan kemudian model dilatih untuk memprediksi kata yang benar pada token yang diberi *mask*, tugas ini dilakukan agar model dapat memahami bahasa dalam level kata dengan lebih baik. Kemudian pada tugas kedua yaitu NSP *input embeddings* terdiri dari dua kalimat yaitu kalimat A dan kalimat B dan model akan dilatih untuk memprediksi apakah kalimat B merupakan lanjutan dari kalimat A, tugas ini dilakukan agar model dapat memahami bahasa pada level kalimat dengan lebih baik.



Gambar 7 Arsitektur *Pra-trained* BERT (Kenton et al., 2018)

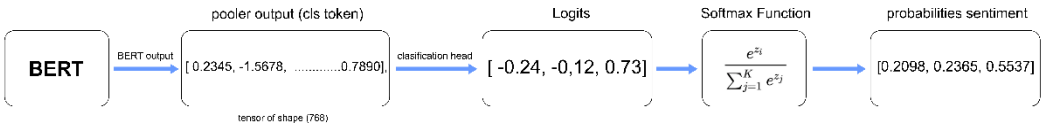
Arsitektur *pre-trained* BERT dapat dilihat pada gambar 7 dimana *input token* terdiri dari dua kalimat yang dipisahkan oleh *token* [SEP] dan salah satu *token* diberi *mask* yang yang ditandai dengan nama TokM, token kemudian di *embeddings* dan diumpan kedalam BERT *neural network* yang terdiri dari *layer encoder transformer*. Pada *output layer* untuk token yang disamarkan BERT akan memprediksi nilai asli dari token tersebut, ini adalah hasil dari tugas MLM. Dan untuk tugas NSP, output dari *token* [CLS] digunakan untuk menjadi representasi vektor yang menentukan apakah kalimat B merupakan lanjutan dari kalimat A. Dengan menjalankan dua tugas tersebut secara bersamaan BERT dapat mempelajari dan memahami bahasa dengan konteks lebih baik.

Pada tahap kedua, dilakukan *fine tune* pada *pre-trained* model dengan menambahkan output *layer* berupa *classification layer* untuk mengklasifikasi kelas sentiment dan dengan melatih model menggunakan dataset yang telah dilabel sebelumnya dengan kelas positif, negatif, dan netral.



Gambar 8 Arsitektur fine tune BERT (Kenton et al., 2018)

Dapat dilihat pada gambar 8 merupakan arsitektur *fine tune* BERT untuk tugas sentiment analisis, pada kasus sentiment analisis data input yang dimasukkan dalam bentuk satu kalimat dan dipecah setiap kata menjadi *token* kemudian token di *embeddings* dan diumpan ke BERT *Neural Network* dan pada output *layer* BERT *node* C merupakan representasi vektor dari konteks data yang di *input* yang kemudian akan dimasukkan kedalam *classification layer* untuk menentukan kelas dari data tersebut (Kenton et al., 2018).



Gambar 9. Classification layer pada BERT

Pada gambar 9, terlihat classification layer yang digunakan pada fine-tuning sentiment analysis. Terdapat tiga proses utama, yaitu Pooler, Classification Head, dan Softmax Function.

Pada tahap awal, BERT menghasilkan output berupa embedding kontekstual untuk setiap token dalam urutan input. Output dari BERT ini termasuk embedding untuk token khusus [CLS] yang ditambahkan pada awal urutan input. Embedding dari token [CLS] ini dikenal sebagai "pooler output" dan memiliki dimensi 768. Token [CLS] ini digunakan karena sebelumnya BERT telah dilatih untuk mempelajari konteks input saat melakukan task NSP (Next Sentence Prediction) dan token [CLS] berisi informasi kontekstual input yang sangat cocok untuk digunakan pada classification task.

Pooler output ini kemudian diteruskan ke "classification head", yang terdiri dari lapisan dense (fully connected). Lapisan dense ini mengubah tensor 768 dimensi menjadi vektor yang lebih kecil yang mewakili logit untuk setiap kelas. Logit ini kemudian melewati fungsi softmax untuk dikonversi menjadi probabilitas (Alex, 2023).

Fungsi softmax menghitung eksponensial dari setiap logit, membaginya dengan jumlah dari semua eksponensial, dan menghasilkan distribusi probabilitas di antara kelas-kelas. Hasil akhirnya adalah distribusi probabilitas atas kelas-kelas sentimen, yang menunjukkan tingkat kepercayaan model terhadap masing-masing kelas.

Dengan menggunakan proses ini, model dapat memanfaatkan kemampuan BERT dalam memahami pola bahasa yang kompleks dan menerapkannya pada tugas klasifikasi sentimen. Pendekatan ini memastikan bahwa model dapat menangkap konteks keseluruhan dari teks input dan menghasilkan prediksi yang akurat berdasarkan informasi tersebut (pavan, 2024).

2.7 IndoBERT

IndoBERT adalah model berbasis *transformer* bergaya BERT, tetapi dilatih murni sebagai *masked language model* (MLM) yang dilatih menggunakan *framework* Huggingface, mengikuti konfigurasi standar untuk BERT-Base yang memiliki 12 *hidden layer*, 12 *attention head* dan *feed-forward hidden layer*. (Koto & Baldwin, 2020). Model IndoBERT yang digunakan pada penelitian ini adalah model yang dikembangkan oleh Sarah Lintang pada tahun 2020 menggunakan arsitektur BERT-Base pada 16 GB data teks mentah yang memiliki lebih dari 2 milyar kata, model ini terbukti dapat mengalahkan performa model multilingual BERT pada tugas *downstream* (Sariwening, 2020).

2.8 Multiclass Confusion Matrix

Confusion matrix adalah matriks yang digunakan untuk melakukan evaluasi proses model klasifikasi berupa jumlah data uji yang benar dan salah. Dengan adanya matriks ini dapat mengetahui kualitas kinerja model klasifikasi (Normawati & Prayogi, 2021).

Matriks ini berisi data target prediksi yang dibandingkan dengan data target aktual. Data prediksi merupakan nilai yang didapatkan dari hasil pemodelan *machine learning*, sedangkan data aktual adalah nilai sebenarnya yang dimiliki. Adanya *confusion matrix* untuk mengetahui sejauh mana *machine learning* bekerja sesuai dengan yang diinginkan. *Confusion matrix* berisi berbagai performa yang dapat diukur seperti akurasi, presisi, *recall* dan F1 Score untuk mengetahui seberapa baik kinerja dari pemodelan yang telah dilakukan sebelumnya (Irwansyah Saputra, 2022).

		PREDICTED classification			
	Classes	a	b	c	d
ACTUAL classification	a	TN	FP	TN	TN
	b	FN	TP	FN	FN
	c	TN	FP	TN	TN
	d	TN	FP	TN	TN

Gambar 10 Multiclass Confusion Matrix

Multiclass Confusion Matrix memungkinkan matriks memvisualisasi kinerja algoritma dalam skenario yang melibatkan lebih dari dua kelas. Setiap baris dari matriks mewakili kelas data aktual, sementara setiap kolom mewakili kelas data prediksi atau sebaliknya (Haghighi et al., 2018).

Precision adalah hasil bagi antara jumlah elemen *True Positive* dan total jumlah unit yang diprediksi sebagai positif (jumlah kolom dari prediksi positif). Lebih detailnya, *True Positive* adalah elemen yang diberi label positif oleh model dan memang positif, sedangkan *False Positive* adalah elemen yang diberi label positif oleh model tapi sebenarnya negatif. *Precision* memberitahu kita seberapa banyak kita dapat mempercayai model ketika memprediksi unit sebagai positif seperti yang ditunjukkan pada formula (1).

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Recall adalah hasil bagi antara jumlah elemen *True Positive* dan total jumlah unit yang diklasifikasi positif (jumlah kolom dari nilai yang benar-benar positif). Lebih detailnya, *False Negative* adalah elemen yang diberi label negatif oleh model tapi nilai sebenarnya adalah positif. *Recall* mengukur akurasi prediksi model untuk kelas positif, ini mengukur kemampuan model untuk menemukan semua unit positif didalam dataset seperti pada formula (2).

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

F1-Score mengevaluasi performa model klasifikasi berdasarkan matriks konfusi. Formula *F1-Score* dapat diinterpretasikan sebagai rata-rata antara *Precision* dan *Recall* yang ditunjukkan oleh formula (3), dimana nilai *F1-Score* mencapai puncaknya pada 1 dan mencapai terendanya pada 0. Kontribusi relative dari presisi dan *recall* adalah setara dalam *F1-Score*, dan rata-rata harmonik digunakan untuk mencari keseimbangan terbaik antara dua ukuran (Grandini et al., 2020).

$$F1\text{-}Score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \quad (3)$$

2.9 Batch Size

Batch size merupakan *hyper-parameter* pada pelatihan model *machine learning* yang menentukan ukuran sebenarnya dari data acak yang dilatih pada setiap *steps* pelatihan (Liu et al., 2019).

Batch size adalah salah satu *hyper-parameter* terpenting yang harus disesuaikan, di setiap epoch, ini adalah jumlah total gambar yang digunakan dalam pelatihan jaringan. Model mungkin memerlukan waktu terlalu lama untuk mencapai konvergensi jika *hyperparameter* disetel terlalu tinggi. Namun, jika disetel terlalu rendah, model akan terpental tanpa mencapai performa yang diinginkan (Aldin & Aldin, 2022).

2.10 Training & Validation Loss

Training & Validation loss merupakan dua metrik penting dalam *machine learning* dan *deep learning* yang digunakan untuk menilai performa suatu model. *Training loss* adalah metrik yang mengukur seberapa cocok model dengan data pelatihan. Ini menilai kesalahan model pada set pelatihan, yang merupakan bagian dari set data yang digunakan untuk melatih model. *Training loss* dihitung dengan mengambil jumlah kesalahan untuk setiap contoh dalam set pelatihan dan biasanya diukur setelah setiap *batch*. Hal ini membantu dalam memahami seberapa baik model mempelajari dan menyesuaikan dengan data pelatihan.

Sebaliknya, *validation loss* digunakan untuk menilai performa model pada set validasi, yang merupakan bagian terpisah dari kumpulan data yang disisihkan untuk memvalidasi performa model. Seperti *training loss*, ini dihitung dari jumlah kesalahan untuk setiap contoh dalam set validasi. *Validation loss* diukur setiap *epoch*, memberikan wawasan apakah model memerlukan penyetelan atau penyesuaian lebih lanjut (Baeldung, 2023).

Training & validation loss sering divisualisasikan bersama dalam grafik untuk mendiagnosis performa model dan mengidentifikasi aspek mana yang perlu disesuaikan. Misalnya, jika *training & validation loss* tinggi, hal ini mungkin mengindikasikan *underfitting*, yang berarti model tidak belajar secara memadai dari

data pelatihan. Jika *training loss* menurun namun *validation loss* meningkat, hal ini dapat mengindikasikan *overfitting*, yaitu model mempelajari data pelatihan dengan terlalu baik namun gagal melakukan generalisasi ke data baru.

Dalam *machine learning*, tujuannya biasanya adalah untuk meminimalkan *training & validation loss*, memastikan bahwa model tidak hanya belajar dengan baik dari data pelatihan tetapi juga menggeneralisasi dengan baik data baru yang belum terlihat (Baeldung, 2023).

Pada pelatihan model BERT untuk tugas analisis sentiment *multi-class*, *Cross-Entropy Loss* biasa digunakan untuk menghitung *training & validation loss* dengan membandingkan distribusi probabilitas model yang diprediksi di seluruh kelas dengan distribusi sebenarnya (label sebenarnya). Adapun cara untuk menghitung *loss* menggunakan *Cross-Entropy Loss* dapat dilihat pada formula (4)

$$Cross - Entropy =$$

$$H = -\sum p(x)\log p(x) \tag{4}$$

Formula diatas menjumlahkan produk dari distribusi probabilitas sebenarnya $p(x)$ untuk sebuah kelas x dan logaritma probabiltas yang diprediksi $q(x)$ untuk kelas yang sama, disemua kelas. Fungsi ini menghitung perbedaan antara probabilitas yang diprediksi dan label kelas sebenarnya. (Vlastimil Martinek, 2020).

2.11 FastAPI

FastAPI adalah framework Python terbaru yang dirancang untuk membangun API yang dapat memberikan kinerja cepat dan efisien, memudahkan pengembangan API yang dapat menghasilkan dokumentasi interaktif dan ramah pengguna. (Chen, 2023) FastAPI dibangun di atas server web Starlette dan menyertakan fitur yang mempermudah pembuatan aplikasi web, seperti validasi data otomatis, penanganan kesalahan, dan dokumen API interaktif. Beberapa kelebihan FastAPI yaitu :

1. Cepat : Menawarkan kinerja sangat tinggi, setara dengan **NodeJS** dan **Go** , berkat Starlette dan pydantic.

2. Cepat dalam mengkode : Memungkinkan peningkatan kecepatan pengembangan yang signifikan.
3. Mengurangi jumlah bug : Mengurangi kemungkinan kesalahan yang disebabkan oleh manusia.
4. Intuitif : Menawarkan dukungan editor yang hebat, dengan penyelesaian di mana saja dan waktu debugging yang lebih sedikit.
5. Mudah : Ini dirancang agar tidak rumit untuk digunakan dan dipelajari, sehingga Anda dapat menghabiskan lebih sedikit waktu untuk membaca dokumentasi.
6. Pendek : Ini meminimalkan duplikasi kode.
7. Kuat : Menyediakan kode siap produksi dengan dokumentasi interaktif otomatis.
8. Berbasis standar : Ini didasarkan pada standar terbuka untuk API, *OpenAPI, dan Skema JSON* .