**Analisis Sentiment Pada Ulasan Vision+ Di Google Play Store Menggunakan Metode Algoritma Bidirectional Encoder Representations from Transformers (BERT)**

**BAB 1: PENDAHULUAN**

* 1. **Latar Belakang**

Dalam era digital yang semakin berkembang pesat, penggunaan teknologi untuk memenuhi kebutuhan hiburan semakin meningkat. Salah satu bentuk hiburan yang sedang populer saat ini adalah platform streaming, yang memungkinkan pengguna untuk menonton konten video secara online. Beberapa platform streaming yang terkenal di Indonesia antara lain Netflix, iFlix, Vidio, dan Vision+.

Vision+ merupakan salah satu platform streaming yang saat ini sedang berkembang pesat di Indonesia. Vision+ menawarkan berbagai jenis konten, seperti film, serial televisi, program acara, dan dokumenter. Namun, dengan semakin banyaknya pilihan platform streaming, persaingan antarplatform semakin ketat. Untuk mempertahankan eksistensinya, platform streaming perlu memperhatikan kualitas konten dan pengalaman pengguna yang disajikan.

Salah satu aspek penting dalam menilai kualitas pengalaman pengguna adalah analisis sentimen. Analisis sentimen merupakan teknik untuk mengekstrak dan memahami perasaan dan opini dari teks yang dituliskan oleh pengguna di media sosial, forum, atau platform online lainnya. Analisis sentimen dapat membantu platform streaming untuk memahami kebutuhan dan preferensi pengguna, sehingga dapat meningkatkan kualitas konten dan pengalaman pengguna.

Dalam konteks Vision+, analisis sentimen berbasis aspek dapat membantu platform untuk memahami bagaimana pengguna merespons konten yang disajikan. Analisis sentimen berbasis aspek merupakan teknik analisis sentimen yang memeriksa perasaan dan opini pengguna terhadap aspek tertentu dari konten dan juga user eksperien dari pengguna platform streaming. Misalnya, dalam konteks film, aspek yang diperiksa dapat meliputi akting, alur cerita, efek visual, kemudahan dalam navigasi aplikasi, kestabilan dari aplikasi dan sebagainya .

Oleh karena itu, penelitian ini akan fokus pada analisis sentimen berbasis aspek terhadap konten dan eksperien penggunaan dari aplikasi platform streaming Vision+. Tujuan dari penelitian ini adalah untuk memberikan pemahaman yang lebih dalam mengenai respon pengguna terhadap aspek-aspek tertentu dari konten dan user eksperien dari penggunaan platform streaming Vision+. Dengan demikian, platform streaming dapat meningkatkan kualitas konten dan pengalaman pengguna dan meningkatkan fitur yang ada di vision+ sehingga dapat mempertahankan eksistensinya di tengah persaingan yang semakin ketat di industri platform streaming.

* 1. **Rumusan Masalah**

Berdasarkan tujuan penelitian yang telah disebutkan sebelumnya, maka rumusan masalah yang akan dijawab dalam penelitian ini adalah sebagai berikut:

1. Bagaimana melakukan analisis sentimen terhadap konten Vision+ menggunakan Bidirectional Encoder Representations from Transformers (BERT)?
2. Bagaimana respon pengguna terhadap aspek-aspek tertentu dari konten Vision+ dalam analisis sentimen berbasis aspek?
3. Apa saja faktor-faktor yang mempengaruhi respon pengguna terhadap konten Vision+ dalam analisis sentimen berbasis aspek?
4. Bagaimana penerapan analisis sentimen berbasis aspek dapat membantu meningkatkan kualitas konten dan pengalaman pengguna di Vision+?

Dengan menjawab rumusan masalah di atas, diharapkan penelitian ini dapat memberikan wawasan dan pemahaman yang lebih mendalam tentang analisis sentimen berbasis aspek dalam konteks platform streaming, khususnya Vision+.

* 1. **Tujuan Penulisan**

Tujuan dari penelitian analisis sentimen adalah sebagai berikut:

1. Untuk melakukan analisis sentimen terhadap konten Vision+ dengan menggunakan metode algoritma Bidirectional Encoder Representations from Transformers (BERT).
2. Untuk memahami respon pengguna terhadap respon positive dan negative dari konten Vision+.
3. Untuk mengukur tingkat akurasi metode algoritma Bidirectional Encoder Representations from Transformers (BERT).
4. Untuk mengevaluasi penerapan analisis sentimen dalam meningkatkan kualitas konten dan pengalaman pengguna di Vision+.

Dengan tujuan tersebut, penelitian ini diharapkan dapat memberikan kontribusi dalam meningkatkan pemahaman tentang penggunaan teknik analisis sentimen berbasis aspek untuk meningkatkan kualitas konten dan pengalaman pengguna di platform streaming. Selain itu, penelitian ini juga diharapkan dapat memberikan rekomendasi dan masukan bagi platform Vision+ dalam mengoptimalkan konten dan pengalaman pengguna.

* 1. **Manfaat Penelitian**

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Manfaat akademis, yaitu sebagai referensi atau sumber penelitian bagi peneliti lain yang tertarik dalam bidang analisis sentimen, khususnya dalam konteks platform streaming.
2. Manfaat praktis, yaitu sebagai masukan atau rekomendasi bagi platform Vision+ dalam meningkatkan kualitas konten dan pengalaman pengguna.
3. Manfaat bagi pengguna platform Vision+, yaitu dalam meningkatkan kualitas layanan dan pengalaman pengguna dalam menikmati konten di Vision+.
4. Manfaat bagi industri platform streaming, yaitu sebagai wawasan dalam menerapkan teknik analisis sentimen berbasis aspek untuk meningkatkan kualitas layanan dan pengalaman pengguna di platform streaming.
5. Manfaat dari tim internal Vision+ yaitu bisa dijadikan masukan kepada tim product dan engineering untuk melakukan improvement dari aplikasi vision+ agar lebih berkembang dan fitur aplikasi lebih relevan dengan kebutuhan user saat ini
   1. **Batasan Penelitian**

Pada subbab ini peneliti menyatakan Batasan serta ruang lingkup yang menjadi titik fokus dari penelitian tugas akhir yang dilakukan peneliti. Yaitu diantaranya;

1. Mengklasifikasikan ulasan positif. netral dan negatif.
2. Penelitian ini menggunakan data ulasan dari google play yang menggunakan bahasa Indonesia
3. Data yang diambil menggunakan ulasan dari pengguna aplikasi Vision+ yang didapat dengan menggunakan package library google play scraper pada Python
4. Penelitian ini menggunakan data yang diambil dari ulasan aplikasi Vision+ dengan Teknik crawling data mencapai 8000 ulasan
   1. **Metodologi Penelitian**

Pada subbab ini peneliti menggunakan python sebagai alat utama dalam mengelola analisis sentiment berikut ini adalah tahapanya:

1. Koleksi Data

Data yang digunakan dalam penelitian ini akan dikumpulkan dari review google play store dari platform streaming Vision+ menggunakan teknik web scraping. Python memiliki beberapa library web scraping yang dapat digunakan, seperti BeautifulSoup , Scrapy dan Google play scraper.Untuk penelitian ini penulis menggunakan Google play scraper untuk memperoleh data mentah.

1. Pra-Pemrosesan Data

Data yang telah dikumpulkan akan melalui proses pra-pemrosesan, seperti menghilangkan duplikasi data, membersihkan data dari karakter-karakter yang tidak diperlukan, serta memfilter data yang tidak relevan dengan penelitian.

1. Analisis Sentiment

Untuk melakukan analisis sentimen, penelitian ini akan menggunakan teknik Natural Language Processing (NLP) dengan Python. Terdapat beberapa library NLP yang dapat digunakan, seperti NLTK dan SpaCy. Teknik ini akan digunakan untuk memisahkan kalimat-kalimat dalam data ke dalam token-token dan kemudian mengklasifikasikan setiap token ke dalam aspek yang sesuai.

1. Pembangunan Model Analisis Sentimen

Model analisis sentimen akan dibangun dengan menggunakan algoritma Machine Learning. Python memiliki banyak library Machine Learning yang dapat digunakan, seperti Scikit-Learn dan TensorFlow. Model ini akan dilatih dengan menggunakan data yang telah dikumpulkan dan telah melalui proses pra-pemrosesan, serta dilakukan validasi dengan menggunakan teknik k-fold cross-validation.

1. Evaluasi Model dan Visualisasi Hasil

Hasil dari analisis sentimen berbasis aspek akan dievaluasi dengan menggunakan beberapa metrik, seperti akurasi, presisi, recall, dan F1-score. Selain itu, hasil dari analisis akan divisualisasikan menggunakan library plotting Python, seperti Matplotlib dan Seaborn, untuk mempermudah pemahaman hasil oleh pembaca.

* 1. **Sistematika Penulisan**

Sistematika penulisan dijelaskan secara garis besar mengenai apa yang terdapat

pada ball-bab yang ada. Adapun sistematika penulisan terdiri dari

1. BAB 1 PENDAHULUAN

Bab ini berisi uraian latar belakang masalah, rumusan masalah, Batasan masalah, tujuan penelitian, metode dan sistematika penulisan.

1. BAB 2 TINJAUAN PUSTAKA

Bab ini berisi tentang teori-teori yang berkaitan dengan penelitian, seperti Analisis Sentimen, Neural Network, Bidirectional Encoder Represenlolions

from Transformers (BERT).

1. BAB 3 METODE PENELITIAN

Bab ini berisi tentang gambaran umum penelitian, sumber data dan tahapan-tahapan yang dilakukan dalam penelitian.

1. BAB 4 IMPLEMEMASI

Bab ini berisi tentang dari hasil pengolahan data. pengujian data, Serta hasil klasifikasi yang telah dilakukan.

BAB 5 PENUTUP

Bab ini berisi kesimpulan dari seluruh hasil penelitian yang merupakan jawaban

dari tujuan dan saran untuk pengembangan yang lebih baik di masa

**BAB 2: Tinjauan Pustaka**

* 1. **Crawling Data**

Crawling Data adalah sebuah proses pengambilan data yang telah tersedia secara online dan terbuka untuk umum yang bertujuan untuk mengumpulkan atau mengunduh data dari suatu database yang bertujuan untuk mengumpulkan atau mengunduh data dari suatu database (Eka Sembodo et al., 2016).

Proses crawling data ini menghasilkan sebuah informasi atau data yang telah kita cari pada worldwide web dan menyimpannya ke dalam file lokal. Data Crawling adalah program yang mengubungkan halaman web, kemudian mengunduh kontennya. Pada penerapannya Crawling Data digunakan untuk mencari suatu hal yang dibutuhkan, salah satunya mencari suatu data dengan penjelajahan target yang lebih luas.

* 1. **Aplikasi Streaming**

Aplikasi streaming adalah program atau aplikasi yang memungkinkan pengguna untuk menonton atau mendengarkan konten multimedia seperti film, acara televisi, musik, dan podcast secara online melalui internet tanpa harus mengunduh terlebih dahulu. Aplikasi streaming ini menggunakan teknologi streaming untuk memutar konten multimedia secara langsung dan memungkinkan pengguna untuk menikmati konten tersebut dalam waktu nyaris real-time.

Menurut Novita (2019), aplikasi streaming adalah suatu jenis aplikasi yang memungkinkan pengguna untuk menonton atau mendengarkan konten multimedia secara online dengan bantuan koneksi internet. Konten multimedia tersebut dapat berupa film, acara televisi, musik, dan podcast yang dapat diputar dengan bantuan teknologi streaming. Aplikasi streaming ini sangat populer dan banyak digunakan oleh masyarakat di seluruh dunia karena memudahkan mereka untuk menikmati konten multimedia tanpa harus membeli atau mendownload konten tersebut terlebih dahulu.

Selain itu, menurut Rianto (2018), aplikasi streaming juga memungkinkan pengguna untuk mengakses konten multimedia dari berbagai perangkat seperti smartphone, tablet, laptop, atau komputer, sehingga memudahkan mereka untuk menikmati konten multimedia kapan saja dan di mana saja.

Dapat disimpulkan bahwa aplikasi streaming merupakan suatu aplikasi yang memudahkan pengguna untuk menonton atau mendengarkan konten multimedia secara online melalui internet dengan bantuan teknologi streaming. Aplikasi ini sangat populer dan banyak digunakan oleh masyarakat di seluruh dunia karena kemudahannya dalam mengakses konten multimedia dari berbagai perangkat dan waktu yang fleksibel.

* 1. **Vision+**

Vision+ adalah platform layanan streaming video on-demand (VOD) yang dikembangkan oleh MNC Group, salah satu perusahaan media terkemuka di Indonesia. Aplikasi ini menyediakan akses ke berbagai konten video seperti film, serial TV, dokumenter, dan program hiburan dari MNC Group dan mitra bisnisnya.

Dalam aplikasi Vision+, pengguna dapat menonton konten video kapan saja dan di mana saja dengan menggunakan perangkat mobile atau tablet yang terhubung ke internet. Aplikasi ini juga menyediakan fitur-fitur seperti fitur pencarian, kategori, rekomendasi, dan nonton offline (download). Aplikasi Vision+ tersedia untuk diunduh secara gratis di Google Play Store dan App Store. Namun, untuk menikmati layanan streaming video, pengguna harus berlangganan layanan berbayar yang disediakan oleh Vision+.

* 1. **Information Technology**

(Brian K. Williams & Stacey C. Sawyer., 2007) mengemukakan pendapat bahwa, teknologi informasi merupakan suatu teknologi yang merupakan hasil gabungan dari jalur komunikasi dengan kecepatan tinggi dengan teknologi komputasi. Menurut pendapat Martin(1999), teknologi informasi tidak hanya memiliki keterbatasan terhadap teknologi komputer yang meliputi perangkat lunak yang biasa disebut software dan perangkat keras atau yang bisa disebut hardware yang digunakan sebagai alat untuk menyimpan serta memproses informasi, melainkan juga mencangkum suatu teknlogi komunikasi yang digunakan untuk mengirimkan suatu informasi ataupun menerima informasi

* 1. **Text Mining**

Text mining merupakan proses menambang data yang berupa teks dimana sumber data secara umum biasanya didapat dari suatu dokumen serta tujuannya adalah mencari kata yang dapat mewakili isi dari suatu dokumen tersebut sehingga dapat dilakukan sebuah analisis yang memiliki keterhubungan antar dokumen tersebut (Zaki Izzani Akbar, 2021). Menurut (Ronen & James, 2007)

Text mining merupakan metode penambangan yang dilakukan oleh komputer dengan tujuan mendapatkan ilmu atau hal baru yang tidak diketahui sebelumnya serta menemukan Kembali informasi yang telah tersebar yang secara otomatis terekstrak dari sumber sumber data teks yang berbeda-beda Text mining merupakan salah satu Teknik yang dapat digunakan untuk melakukan klasifikasi. Teknik ini menemukan pola menarik dari suatu pengumpulan data tekstual (Santi et al., 2020). Text mining mirip dengan data mining. Perbedaannya, Teknik yang dirancang untuk bekerja pada data yang terstruktur atau semi terstruktur.

* 1. **Analisis Sentiment**

Analisis sentiment adalah suatu proses mengestraksi, mengolah, serta memahami suatu data yang berbentuk teks yang tidak terstruktur secara otomatis yang bertujuan untuk mengambil suatu informasi sentiment yang terdapat pada sebuah kalimat pendapat atau opini terhadap suatu topik baik negatif maupun positif (Imam Fahrur Rozi et al., 2012). Metode pemrosesan ini dapat diterapkan pada semua bidang opini yang ada. Diantaranya pada topik ekonomi, politik, sosial ataupun hukum.

*Sentiment analysis* atau analisis sentimen dalam bahasa indonesia adalah sebuah teknik atau cara yang digunakan untuk mengidentifikasi bagaimana sebuah sentimen diekspresikan menggunakan teks dan bagaimana sentimen tersebut bisa dikategorikan sebagai sentimen positif maupun sentimen negative (Muchammad Shiddieqy Hadna et al., 2016). Selain itu (Medhat et al., 2014) mengemukakan bahwa Sentiment Analysis atau Opinion Mining merupakan sebuah studi yang mempelajari mengenai pendapat, sikap, dan emosi seseorang terhadap suatu entitas yang dapat mewakili suatu individu, acara atau suatu topik tertentu. Analisis sentimen mencakup deteksi, analisis, dan evaluasi keadaan pikiran manusia terhadap berbagai peristiwa, masalah, layanan atau minat lainnya (Steven, 2020) Lebih tepatnya, bidang ini bertujuan untuk menggali pendapat, sentimen dan emosi berdasarkan pengamatan orang-orang yang bisa didapatkan melalui tulisan, ekspresi wajah, ucapan, musik, gerakan, dan lain sebagainya (Steven, 2020)

Tujuan dari analisis sentimen sendiri adalah untuk menemukan pendapat, mengidentifikasi sentimen yang mereka ungkapkan, dan kemudian mengklasifikasikan polaritasnya (Medhat et al., 2014). Dengan kata lain, analisis sentimen berfungsi untuk mengklasifikasikan teks kedalam kelas positif, negatif atau netral (Steven, 2020) Beberapa pendapat mengenai analisis sentimen dapat diambil kesimpulan bahwa analisis sentimen adalah suatu proses untuk menentukan sentimen atau opini dari seseorang yang biasanya di wujudkan dalam bentuk teks dan bisa dikategorikan sebagai sentimen positif atau negatif (Muchammad Shiddieqy Hadna et al., 2016)***.***

* 1. **Neural Network**

Neural Network atau jaringan syaraf tiruan adalah salah Satu teknik machine

learning yang vuvuler dengan mensimulasikan mekanisme 'Embelajaran yang

terinspirasi dari bagaimana cara sistem syaraf manusia atau makhluk biologis lainnya.Sistem Saraf terdiri dari sel yang disebut dengan neuron. Neuron-neuron tersebut Saling terhubung dengan satu sama lain axon dan dendrites. Sinapsis adalah penghubung antara axon dan dendriles (Aggarwal, 2018).

Network sendiri disebut sebagai arsitektur, di mana berbagai layers saling

terhubung dengan Satu sama lain. Layer yang ada di antara input layer dan Output layer disebut dengan hidden layers dan output dari hidden layer disebut dengan hidden units

(Osinga, 2018). Istilah hidden diberikan karena unit-unit tersebut tidak dapat langsung terlihat dari luar sebagai input atau output. Inti dari neural network adalah hidden layer yang dibentuk dari hidden units, yang masing-masing merupakan neural units,mengambil weighted sum dari input dan kemudian menerapkan non-lineariw.

Setiap unit pada tiap layer mengambil input dan Output dari semua unit di layer

dan hubungan antara setiap pasangan unit daridua layer yang saling berdekatan

sehingga setiap layer saling terhubung. Tiap hidden units akan menjumlahkan semua input unit (Jurafsky & Martin, 2019).Jaringan syarat tiruan memiliki beberapa parameter, seperti weights W, bias b yang dipelajari Oleh gradient descent. Sedangkan hyperparameters adalah parameter yang dipilih Oleh desainer algoritma nilai optimal yang diatur pada sebuah devser,bukan dengan sebuah gradient descent di dataset training. Hyperparameters mencakup learning rate n, mini-barch size, arsitektur model (jumlah layer, jumlah hidden node tiap layer, fungsi aktivasi yang dipilih) dan lain-lain (Jurafsky & Maltin, 2019).

Gradient descent (penurunan gradien) adalah sebuah algoritma optimasi yang

digunakan ketika melatih sebuah model machine learning. Gradient descent hanya

digunakan untuk menemukan nilai parameter fungsi (koefisien) yang meminimalkan

cost. Tujuan utama dari gradient descent adalah menemukan weight yang optimal

dengan meminimalisasi loss function. Seberapa besar step yang dilakukan gradient

descent ditentukan Oleh learning rare yang menunjukkan seberapa cepat atau lambat

bergeraknya fungsi ke weight yang optimal. Terdapat tiga jenis gradient descent yang

sering digunakan, yaitu Balch Gradient Descent, Stochastic Gradient Descent, dan

Mini -Batch Gradient Descent.

1. Batch gradient descent (BGD) atau vanilla gradient descent menghitung error dari

tiap contoh yang ada pada dataset training. proses ini seperti sebuah siklus yang

disebut juga training epoch. Semua cost dari training pada daraser dihitung.

Algoritma ini memiliki kelebihan yaitu dapat membuat jaringan syaraf tiruan

lebih cepat, menghasilkan gradien kesalahan yang stabil, Akan tetapi

gradien kesalahan yang stabil dapat pula membuat model tidak dapat menghasilkan

kondisi konvergensi yang terbaik\_ Selain itu,seluruh training dataset harus berada

dalam memori.

1. Stochastic gradient descent (SGD) adalah algoritma yang meminimalkan loss

function dengan menghitung gradiennya tiap kali 'raining dilakukan (Jurafsky &

Martin, 2019). Algoritma ini disebut stokastik karena hanya memiliki Satu contoh

acak dalam satu waktu, memindahkan weight-nyasehingga dapat meningkatkan

Irrforma dari contoh tersebut. Sehingga, SGD akan melakukan proses tersebut pada

setiap contoh dataset Satu per Satu. Weight yang terus diperbarui memungkinkan

untuk mendapatkan peningkatan yang cukup detail. Akan tetapi pembahaman yang

terlalu sering dilakukan lebih mahal dan dapat menyebabkan gradient yang tidak rata karena error rare dapat naik turun.

1. Mini barch gradient descent menggabungkan konsep dari SGD dan BGD. Mini

batch adalah pruses training sekumpulan dataset (umumnya512 atau 1024) yang

lebih sedikit dari dataset asli. Algoritma ini akanmembagi training dataset menjadi

batch-barch yang kecil dan memperbarui weight dari masing-masing batch.

Salah Satu algoritma optimasi yang sering digunakan adalah Adam. Adam atau

Adaptive Momenl Eslimarion Algorithm (Kingma & Ba, 2015) menghitung estimasi

momen dan menggunakannya untuk mengoptimmsi fungsi. Adam merupakan

kombinasi dari dua algoritma yaitu AdaGrad dan RMSProp. AdaGrad akan

mempertahankan learning rare per parameter yang meningkatkan peforma pada

masalah dengan gradien yang renggan dan RMSProp juga mempertahankan learning

rare per parameter yang diadaptasi berdasarkan rata-rata besaran gradien untuk weight

(s&rapa cepat berubah). Algoritma ini akan menghitung rata-rata eks1Nmensial

weighted bergerak kemudian mengkuadratkan gradien yang dihitung. Alih•alih

mengadaptasi learning rate parameter berdasarkan rata-rata momen Ikrtama seperti di

RMS Prop, Adam juga menggunakan rata-rata momen kedua. Dengan menggunakan

model dan dataset yang besar, Adam dapat secara efektif menyelesaikan masalah•

masalah deep learning praktikal (Kingma & Ba, 2015). Hal ini membuat Adam

optimizer menjadi metode yang efisien secara komputasi, membutuhkan sedikit

memori, tidak berbeda dengan gradien dengan skala diagonal, cocok untuk masalah

yang menggunakan data atau parameter yang besar (Kingma & Ba, 2015).

* 1. **Artificial Neural Network ANN**

Artificial neural network (ANN) atau jaringan saraf tiruan adalah model matematis yang terinspirasi dari sistem saraf biologis. ANN terdiri dari jaringan neuron buatan (artificial neuron) yang terhubung secara paralel, dan mampu melakukan pemrosesan informasi melalui propagasi sinyal yang disebut "forward pass" dan "backward pass".

Setiap neuron dalam ANN menerima masukan (input) dari neuron lain atau dari luar, dan kemudian menghasilkan keluaran (output) berdasarkan suatu fungsi aktivasi yang dapat berbeda-beda, seperti fungsi sigmoid atau fungsi ReLU. Hubungan antara neuron-neuron dalam ANN dapat direpresentasikan oleh bobot (weight), yaitu nilai numerik yang menentukan seberapa besar pengaruh suatu neuron terhadap neuron lainnya.

Proses pembelajaran (learning) ANN terjadi melalui pengaturan bobot dengan tujuan untuk meminimalkan suatu fungsi error. Terdapat berbagai macam algoritma pembelajaran ANN, seperti backpropagation, gradient descent, dan stochastic gradient descent, yang digunakan untuk menentukan perubahan bobot yang optimal.

ANN digunakan dalam berbagai aplikasi seperti klasifikasi, prediksi, pengenalan pola, dan optimasi. Contoh penggunaan ANN antara lain dalam pengenalan wajah, deteksi objek, dan prediksi harga saham.

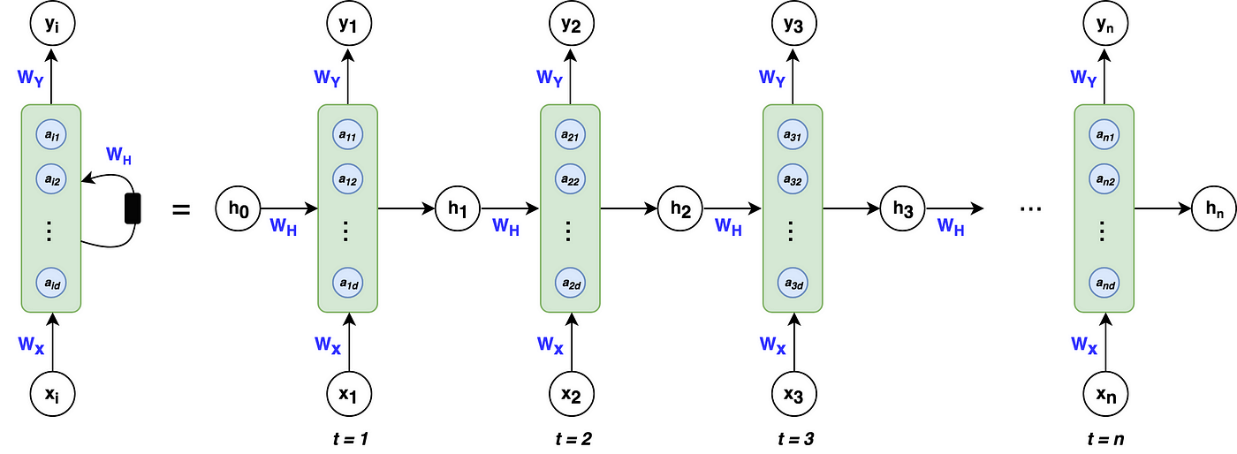
* 1. **Recurrent Neural Network**

Recurrent neural network (RNN) atau jaringan saraf rekuren adalah model ANN yang dirancang untuk mengolah data sekuensial, seperti teks atau waktu. RNN memungkinkan informasi dari masa lalu untuk mempengaruhi pengambilan keputusan di masa sekarang melalui penggunaan sel memori (memory cell) dan loop feedback (kembali).

RNN terdiri dari serangkaian sel memori (memory cell) yang terhubung ke neuron-neuron lainnya. Setiap sel memori menghasilkan output berdasarkan input sekarang dan output sebelumnya. Output sebelumnya disimpan dalam sel memori, dan digunakan untuk mempengaruhi pengambilan keputusan selanjutnya.

Salah satu jenis RNN yang populer adalah LSTM (Long Short-Term Memory) yang memungkinkan sel memori untuk menyimpan informasi dalam jangka panjang dan jangka pendek. Hal ini memungkinkan LSTM untuk lebih efektif mengolah data sekuensial, terutama pada kasus di mana ada jarak waktu yang cukup jauh antara suatu kejadian dan kejadian berikutnya.

Penggunaan RNN terutama berfokus pada pengolahan data sekuensial seperti teks, waktu, dan suara. Beberapa aplikasi dari RNN antara lain terjemahan mesin, pengenalan suara, dan analisis sentimen.Berikut ini adalah gambaran dari Recurrent Neural Network (RNN)



Gambar 2.1 Recurrent Neural Network

* 1. **Machine Learning**

Machine learning (ML) adalah bidang ilmu komputer yang mempelajari algoritma untuk membuat mesin dapat belajar dari data, tanpa perlu diprogram secara eksplisit. Dalam ML, mesin belajar dengan mengidentifikasi pola pada data yang diberikan, dan kemudian menggunakannya untuk membuat prediksi atau mengambil tindakan.

Ada tiga jenis utama dari machine learning:

1. Supervised learning, di mana model belajar dari data yang diberi label dan melakukan prediksi untuk data baru.
2. Unsupervised learning, di mana model belajar dari data tanpa label dan mencoba untuk mengidentifikasi pola atau klasifikasi pada data tersebut.
3. Reinforcement learning, di mana model belajar melalui interaksi dengan lingkungannya dan mendapatkan reward atau hukuman berdasarkan tindakan yang diambil.

ML telah digunakan secara luas dalam berbagai aplikasi, termasuk pengenalan wajah, klasifikasi teks, deteksi spam, dan pengenalan suara. Peningkatan teknologi komputer dan perkembangan algoritma ML, seperti deep learning, telah memungkinkan mesin untuk menyelesaikan tugas yang lebih kompleks dan lebih presisi.

* 1. **Deep Learning**

Deep learning merupakan cabang dari machine learning yang masih menjadi

bagian dari kecerdasan buatan. Deep learning merupakan neural network yang lebih modern dan bersifat deep atau mendalam karena memiliki jauh lebih banyak layer dibandingkan dengan dengan neural network biasanya (Jurafsky & Martin, 2010) (Chollet. 2018). Deep learning bekerja untuk sehingga tidak hanya dapat memprediksi tetapi juga merepresentasikan data dengan benar, sehingga cocok untuk melakukan prediksi (Goldberg, 2017). Deep learning dapat dibagi kedalam tiga metode pendekatan yaitu supervised, semi-supervised dan unsupervised learning. Deep learning didukung Oleh banyak framework seperti TensorFlow, Torch dan Theano.

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

Bidirectional Encoder Representations from Transformers atau disingkat BERT

adalah suatu model representasi bahasa terlatih yang dikembangkan Oleh para peneliti di Google AI Language pada tabun 2018. BERT merupakan sebuah teknik open source yang dikembangkan berdasarkan teknik deep learning dan berbagai metode seperti semi-supervised learning, ELMO, OpenAI Transformer dan Transformer. Sesuai dengan namanya, BERT menggunakan Transformer. Transformer adalah mekanisme yang hubungan kontekstual antara kata-kata dalam teks (Vaswani et al. 2017). Transformer dapat memahami dan mengkonversi yang diperoleh dengan mekanisme yang bernama self-attention mechanism. Self-arrention mechanism adalah cara Transformer untuk mengubah "pemahaman" kata terkait lainnya menjadi kata-kata yang akan diproses dengan mekanismenya. Terdapat dua mekanisme pada Transformer, yaitu:

1. Encoder

Encoder memiliki fungsi untuk membaca keseluruhan teks inputan sekaligus.

encoder terdiri dari stack N = 6 layers yang identik. Setiap layer memiliki dua sub

layer yaitu self-attention layer dan feed-forward neural network. Dengan self-

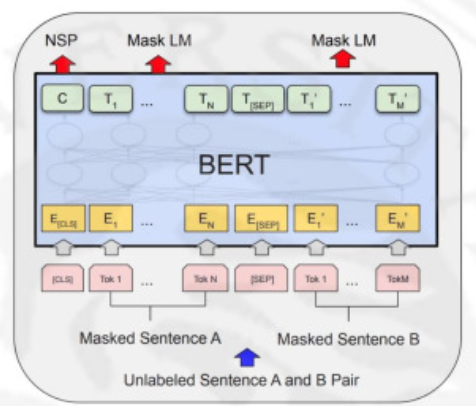
attention layer, encoder dapat membantu node untuk tidak hanya fokus kepada

kata yang sedang dilihat tetapi juga untuk mendapatkan konteks semantik dari

kata tersebut.

1. Decoder

Decoder berfungsi untuk menghasilkan urutan output yang benlpa prediksi.Decoder juga terdiri dari stack N = 6 layers yang identik. Setiap layer terdiri dari dua sub layer Seperti yang ada pada encoder dengan tambahan attention layer diantara dua layers tersebut untuk membantu node Saat ini mendapatkan key content yang membutuhkan attention (Vaswani et al. 2017) dengan melakukan multi-head attention pada output dari encoder. Sama dengan di encoder, self-attention layer di encoder membuat setiap posisi di decoder dapat menangani semua sebelumnya dan posisi saat itu.



Gambar 2.2. Arsitektur Model BERT

BERT di-training dengan dataset, dataset harus disesuaikan dengan representasi input yang dapat diterima Oleh BERT. Oleh karena itu dibutuhkan sebuah tokenizer yang lx-rtujuan untuk melakukan tokenisasi pada kalimat-kalimat dan menghasilkan input yang sesuai. Hal ini dilakukan karena BERT menggunakan vocabulary yang spesifik yang mana tergantung dengan model apa yang dipakai.Vocabulary dibuat dengan menggunakan sebuah model WordPiece. Vocabulary terdiri

dari semua karakter dan tambahan sekitar 30 kata-kata yang sering digunakan Serta

sub kata yang sering muncul pada corpus yang model latih. BERT menerima panjang yang tetap dan sama untuk setiap inputnya. BERT telah menentukan panjang urutan kalimat maksimumnya, yaitu 512 karena encoder pada Transformer hanya menghasilkan output dengan dimensi 512 saja. Jika panjang kalimat lebih dari panjang maksimum yang telah ditentukan, kalimat akan dikurangi (truncate).Sedangkan jika panjang urutan kalimat kurang dari panjang maksimum yang sudah ditentukan, kalimat akan ditambah dengan padding. Akan tetapi, panjang maksimum juga dapat ditentukan sendiri tergantung dengan kebutuhan dan tugas yang akan dilakukan. Proses mempersiapkan kalimat menjadi representasi input pada BERT dilakukan Oleh tokenizer, dengan langkah-langkah sebagai berikut:

1. Setiap kalimat ditokenisasi menjadi per kata atau sub kata menggunakan

Wordpiece. Untuk melakukan tokenisasi pada sebuah kata, tokenizer akan

memeriksa apakah tiap kata pada kalimat terdapat pada vocabulary. Jika tidakada,

tokenizer akan memecah kata menjadi sub-sub kata yang kemungkinan

kemunculannya pada vocabulary paling be.sar. Jika tokenizer juga tidak menemukan

sub kata pada vocabulary, kata tersebut dilrcah menjadi per karakternya. Akan

tetapi, jika semua kata diubah menjadi sub kata atau karakter individual, akan terjadi

overload Kata-kata yang tidak ada pada vocabulary akan diganti dengan token

[UNKI atau unknown. Namun jika semua kata diubah menjadi token tersebut,

banyak informasi yang akan hilang\_ Oleh kalena itu, kata-kata dapat dipecah

menjadi sub kata dengan simbol BERT melakukan ini karena dua hal yaitu

pertama untuk mem1Rrcepat processing dan mengurangi jumlah parameters yang

harus dilatih, dan kedua untuk mengatasi masalah out-of-vocabulary.

1. Setiap kalimat diberi token-token khusus yaitu [CLSI di awal kalimat dan [SEPI di

akhir kalimat. Token [CIS] menjadi indikator sebuah kalimat sekaligus

sentimennya Saat klasifikasi sentimen dilakukan. Token [SEP]adalah token yang

digunakan untuk memisahkan kalimat satu dengan kalimatselanjutnya. Kalimat

yang sudah diberi token khusus ini menjadi token embeddings.

1. Setelah itu kalimat-kalimat disesuaikan dengan panjang maksimum yang telah

ditentukan dengan mengurangi atau memberi padding dengan token khusus [PAD].

1. Kemudian setiap kalimat dicocokkan dengan bilangan unik atau ID sesuai dengan vocabulary dan bilangan unik tersebut disimpan sebagai token id. Bilangan unik atau ID ini didapatkan ketika model di-rraining, yang mana membuat setiap kata,sub kata, dan karakter pada vocabulary memilikibilangan uniknya masing-masing.ID tersebut diperoleh berdasarkan indeks kata pada vocabulary karena pada vocabulary disusun berdasarkan tingkat kemunculannya. Kata dan sub kata sendiri harus dikonversi menjadi ID-nya karena model pre-trained BERT hanya dapat mengerti ID dari token.
2. Sentence embedding diberi pada kalimat untuk membedakan kalimat pertama dan kalimat kedua atau padding. Tahap ini dapat dilakukan dengan memberi angka 1 pada kalimat lwrtama dan memberi angka 0 pada padding. Tokenizer dapat mengetahui mana kalimat dan mana padding dengan melihat token [SEPI] yang berfungsi memisahkan antara dua kalimat.
3. Positional embedding juga ditambahkan untuk tiap token untuk menunjukkan isi tiap kata pada kalimat\_ Tahap ini dilakukan karena BERT tidak mengetahui posisi dari tiap kata. Sehingga walaupun pada awal kalimat terdapat suatu kata dan di akhir kalimat juga terdapat kata yang sama tetapi memiliki makna yang berbeda, BERT tidak akan memperlakukan token tersebut dengan makna yang sama. BERT akan menerima unltan kata-kata atau sebuah kalimat sebagai input yangakan terus melalui tumpukan encoder. Tiap encoder mengaplikasikan self-attention dan memberikan output melalui feed-forward network yang kemudian dilanjutkan Oleh encoder selanjutnya. Pada penelitian ini, model yang dipilih adalah BERT-base sehingga proses ini berlanjut sebanyak dua belas Setelah melewati semua encoder, tiap token per posisi memberikan output vektor dengan ukuran hidden size yaitu 768 pada BERT-base. Untukproses analisis sentimen, output yang diperhatikan adalah output dari pertama yaitu token [CLSI. Vektor tersebut digunakan sebagai input untuk classifier. BERT dapat mencapai hasil yang baik hanya dengan menggunakan neural network tunggal sebagai classifieFnya (Devlin et al., 2019). Layer yang digunakan untuk klasifikasi adalah fully connected neural network dengan fungsi sofrmax. Sehingga output dari BERT yang digunakan untukklasifikasi berasal vektor token [CLSI] karena token [CIS] dianggap melakukan pengumpulan rata-rata ata.S token kata untuk mendapatkan vektor dari kalimat. Layer terakhir pada classifier layer menghasilkan logits. Logits adalah output yanglR-rupa prediksi probabilitas kasar dari kalimat yang akan diklasifikasikan. Softmax akan mengubah logits tersebut menjadi probabilitas dengan mengambil eksponen dari tiap nilai logit sehingga total probabilitasnya adalah tepat sehingga nilai probabilitas akan berada di antara 0 atau angka positif.
   1. **IndoNLU**

IndoNLU merupakan salah satu produk dari komunitas yang bernama *IndoBenchmark*. IndoNLU *Benchmark* adalah sekumpulan sumber daya untukmelatih, mengevaluasi, dan menganalisis system pemahaman bahasa alami untuk bahasa Indonesia. IndoNLU mencakup dua belas *tasks*, mulai dari klasifikasi kalimat tunggal hingga pelabelan urutan pasangan kalimat dengan tingkat kompleksitas yang berbeda. Kumpulan data untuk *tasks* terletak pada *domain* dan *styles* yang berbeda untuk memastikan keragaman *task*.

* 1. **IndoBERT-BASE**

IndoNLU mengikuti pengaturan *hiperparameter pre-training* BERT untuk melatih model IndoBERT terlebih dahulu. IndoBERT merupakan model berbasis BERT berbahasa Indonesia yang dilatih megggunakan dataset indo4B dengan 4 milyar kata bahasa Indonesia dan 20 GB ukuran data teks. Terdapat dua model IndoBERT berdasarkan ukurannya yaitu IndoBERTBASE dan IndoBERTLARGE. IndoBERT-base memiliki parameter sebesar 124,5M, 12 *layer*, 12 *head*, 768 *embedding size*, 728 *hidden size*, dan tipe *pre-train* dengan metode kontekstual.

* 1. **Confusion Matrix**

*Confusion Matrix* dapat diartikan sebagai suatu alat yang memiliki fungsi untuk melakukan analisis apakah *classifier* tersebut baik dalam mengenali *tuple* dari kelas yang berbeda. nilai dari *True-Positive* dan *True-Negative* memberikan informasi ketika *classifier* dalam melakukan klasifikasi data bernilai benar, sedangkan *False-Positive* dan *False-Negative* memberikan informasi ketika *classifier* salah dalam melakukan klasifikasi data.

Tabel 2.1 *Confusion Matrix*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | Predicted Class |  |  |
|  |  |  |  |  |  |
|  |  | Positive | Neutral | Negative |  |
|  |  |  |  |  |  |
| True | Positive |  |  |  |  |
| Class |  | True Positive | False Neutral | False Negative |  |
|  |  |  |  |  |
|  |  | (TP) | (FNt) | (FN) |  |
|  |  |  |  |  |  |
|  | Neutral | False Positive | True Netral | False Negative |  |
|  |  | (FP) | (TNt) | (FN) |  |
|  |  |  |  |  |  |
|  | Negative |  |  |  |  |
|  |  | False Positive | False Neutral | True Negative |  |
|  |  | (FP) | (FNt) | (TN) |  |
|  |  |  |  |  |  |

Keenam istilah tersebut adalah :

* *True Positive* (TP), merupakan data positif yang diprediksi benar
* *True Negative* (TN), merupakan data negatif yang diprediksi benar
* T*rue Neutral* (TNt) adalah kalimat memiliki sentimen netral dan hasil prediksinya juga menunjukkan sentimen netral.
* F*alse Neutral* (FNt) adalah kalimat yang memiliki sentimen postif atau negatif tetapi hasil prediksinya menunjukkan sentimen entral
* *False Positive* (FP), merupakan data netral atau negatif namun diprediksi sebagai data positif
* *False Negative* (FN), merupakan data positif atau netral namun diprediksi sebagai data negatif

Beberapa hal yang dapat dilakukan oleh *Confusion Matrix* dalam hal perhitungan :

1. *Accuracy*

*Accuracy* memberikan rasio prediksi benar (positif dan negatif) dengan keseluruhan data. nilai *accuracy* diperoleh dengan persamaan berikut :

1. *Precision*

*Precision* memberikan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Nilai *precision* diperoleh dengan persamaan berikut :

1. *Recall*

*Recall* memberikan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Nilai *Recall* diperoleh dengan persamaan berikut :

1. *F1 Score*

*F1 Score* merupakan perbandingan rata-rata presisi dan *recall* yang dibibitkan. Nilai *F1 Score* dapat diperoleh dengan persamaan berikut :

* 1. **Dataset**

Dataset adalah sekumpulan data yang bersifat sebagai himpunan data yang berasal dari informasi-informasi pada masa sebelumnya dan siap untuk di kelola menjadi sebuah informasi baru dengan menggunakan teknik pembelajaran *supervised learning*.Dimana digunakan untuk sistem prediksi sebagai acuan pendukung keputusan (Khalimi, 2020).

* 1. **Web Scrapping**

Web scraping mempakan sebuah metode pengumpulan data melalui internet,

metode web scraping populer diigunakan dalam lymenuhan data mining. Menurut teori, web scraping adalah cara untuk mengumpulkan data menggunakan melode yang berbeda dengan penggunaan API (*Application programming Interfaces*). Cara seperti ini biasanya dimulai dengan penulisna kode program, dimana digunakan sebagai otomatisasi *query* untuk melakukan *request* data terhadap server. Data hasil tersebut dapat dilakukan ekstrasi untuk menghasilkan informasi yang akan dicari (Sahrian. 2020).

* 1. ***Case Folding***

Case folding merupakan suatu bentuk tert preprocessing yang sederhana dan

efektif meskipun sering diabaikan. Case folding memiliki tujuan untuk mengubah

semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf 'a' sampa 'z' yang

diterima. Karakter selain huruf dihilangkan dan dianggap delimiter. Ada beberapa cara yang dapat digunakan dalam tahap case folding, anda dapat menggunakan atau semuanya, tergantung pada tugas yang diberikan (Nugroho, 2019).

* 1. ***Dana Cleaning***

Data Cleaning merupakan tahap pembersihan dengan kata-kata yang tidak perlu sehingga dapat meningkatkan akurasi dari analisis yang akan dilakukan. Proses yang dilakukan dalam tahap cleansing adalah sebagai berikut:

1. Menghapus https
2. Menghapus mention
3. Menghapus hashtags
4. Menghapus tanda baca
5. Menghapus angka
6. Menghapus spasi berlebih
7. Menghapus kalimat duplikat.
   1. **Normalisasi Teks**

Normalisasi teks merupakan proses pengolahan teks yang bertujuan untuk

mengubah struktur atau bentuk teks yang asalnya sulit dimengerti komputer hingga

akhirnya mudah dimengerti dan diolah lebih lanjut. Normalisasi teks meliputi

pengubahan singkatan, akronim, kesalahan penulisan dan penggunaan kata tidak baku.

* 1. **Dataset Splitting**

Dataset splitting merupakan suatu metode yang digunakan untuk mengevaluasi

performa model. Metode ini membagi dataset menjadi beberapa bagian yakni bagian

yang digunakan untuk training data, validation data dan untuk testing data dengan

proporsi tenentu. Dataset training digunakan untuk melatih model, dataset validasi

digunakan untuk meminimalisir overfilling yang sering terjadi pada jaringan syaraf turuan sedangkan dataset testing sendiri digunakan sebagai test akhir untuk melihat keakuratan jaringan yang sudah dilatih dengan dataset training

* 1. **Python**

*Python* adalah bahasa pemrograman tingkat tinggi yang banyak digunakan akhir-akhir ini. dengan filosofi dimana kode mudah dibaca dan sintaksnya mengizinkan *programmer* untuk mengekspresikan konsep dalam kode baris yang lebih sedikit daripada di bahasa pemrograman C. kontruksi bahasa memungkinkan pengguna untuk menulis program secara jelas baik dalam skala kecil maupun besar.

Fitur terpenting dalam *python* adalah mendukung beberapa paradigma pemrograman, termasuk berorientasi objek, pemograman atau prosedural imperatif dan fungsional gaya. *Python* mendukung sistem tipe dinamis dan manajemen memori otomatis dan memiliki besar dan perpustakaan standar yang komprehensif. Penerjemah *python* tersedia untuk banyak sistem operasi.

* 1. **Google Colaboratory**

Google Colab atau Google Colaboratory, adalah sebuah *executable document* yang dapat digunakan untuk menyimpan, menulis, serta membagikanprogram yang telah ditulis melalui Google Drive. *Software* ini pada dasarnya serupa dengan Jupyter Notebook gratis berbentuk *cloud* yang dijalankan menggunakan *browser*, seperti Mozilla Firefox dan Google Chrome. Google Colabmemungkinkan user untuk menjalankan kode Python tanpa perlu melakukan proses instalasi dan setup lainnya karena semua keperluan *setting* dan *adjustment* akan diserahkan ke cloud (Glints.com, 2021).

Google Colaboratory dapat menulis dan mengeksekusi Python di *browser* dengan beberapa keuntungan yaitu tidak memerlukan konfigurasi, akses gratis ke GPU, berbagi dengan mudah serta dapat menggabungkan kode yang dapat dijalankan dan *rich text* dalam satu dokumen, beserta gambar, HTML, LaTeX, dan lainnya (*colab.research.google.com*). Dengan sifat yang fleksibel Google Colab dapat membuat *notebook* Colab oleh *user* itu sendiri, *notebook* tersebut akan disimpan di akun Google Drive pengguna. Pengguna dapat dengan mudah membagikan *notebook* Colab dengan pengguna lainnya, sehingga dapat saling memberikan komentar di *notebook* pengguna satu sama lainnya atau bahkan mengeditnya.

* 1. **Penelitian Terkait**

Penelitian ini dilakukan dengan mempelajari penelitian terdahulu yang memiliki keterkaitan tentang analisis sentimen. Beberapa penelitian terdahulu yang digunakan sebagai dasar pembelajaran pada Tabel 2.1

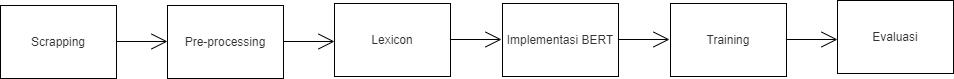
Tabel 2.1 Penelitian Terkait

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Nama | Tahun | Judul | Hasil |
| 1 | Cindy Alifia  Putri,  Adiwijaya,  Said Al Faraby | 2020 | Analisis Sentimen  Review Film  Berbahasa Inggris  dengan Pendekatan  Bidirectional Encoder  Representations from  Transformers | Akurasi yang didapatkan  dari analisis sentimen  menggunakan metode  BERT sebesar 73%.  dataset yang digunakan  berbentuk dokumen.  BERT-base terbukti |
| 2 | Raden Mas  Rizi Wahyu  Panca Kusuma  Atmaja, Wily  Yustanti | 2020 | Analisis Sentimen  Customer Review  Aplikasi Ruang Guru  dengan Metode  BERT(Bidirectional  Encoder  Representations from  Transformers) | Akurasi yang didapatkan  dari analisis sentimen  menggunakan metode  BERT sebesar 99%  dengan nilai F1 Score  sebesar 98.9%.  menggunakan 5437  testing data dimana 5254  komentar positif, 16  netral dan 167 komentar  negatif. |
| 3 | Muhammad  Mahrus Zain,  Rizky  Nathamael  Simbolon,  Harlem  Sulung, Zaidan  Anwar | 2021 | Analisis Sentimen  Pendapat Masyarakat  Mengenai Vaksin  Covid-19 Pada Media  Sosial Twitter dengan  Robustly Optimized  BERT Pretraining  Approach | Akurasi yang didapatkan  dari analisis sentimen  menggunakan metode  RoBERT sebesar 95%.  dimana rata-rata hasil  akurasi prediksi pada  label positif 84%, Netral  97% dan Negatif 93% |

**Bab 3 Metodologi Penelitian**

* 1. **Tahapan Penelitian**

Tahapan penelitian analisis sentimen yang dilakukan melalui komentar di Google Playstore pada aplikasi Vision+ untuk melihat bagaimana sentimen atau pendapat masyarakat mengenai aplikasi tersebut. Analisis sentimen ini menggunakan metode BERT untuk mengklasifikasikan data komentar yang mengandung kata Vision+ ke dalam tiga kategori, yaitu positif, netral, dan negatif. Diagram proses dari analisis sentimen dapat dilihat pada Gambar 3.1.



Gambar 3.1. Tahapan Penelitian

* 1. **Scrapping**

*Scrapping* atau pengambilan data dilakukan dengan metode *scraping* data di platform Google Play Store menggunakan *python script* dengan *library google* *play scraper*. *Scraping* dilakukan untuk mendapatkan komentar pengguna Disney+ Hotstar yang tertulis pada platform Google Play Store seperti pada contoh berikut:

#scrape jumlah ulasan yang diinginkan

from google\_play\_scraper import Sort, reviews

result, continuation\_token = reviews(

'com.zte.iptvclient.android.idmnc',

lang='id', #disini kita mau men scrape data ulasan aplikasi shopee yang berada di google play store

country='id', #kita setting bahasa nya menjadi bahasa indonesia

sort=Sort.MOST\_RELEVANT, # # kemudian kita gunakan most\_relevan untuk mendapatkan ulasan yang paling relevant

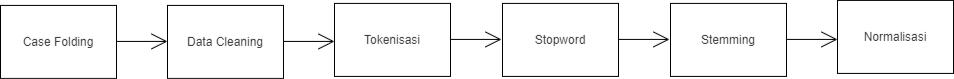
count=8000, # disini jumlah ulasan yang mau kita ambil ada seribu

filter\_score\_with=None # # kemudian di filter\_score kita gunakan None untuk mengambil semua score atau ratting bintang 1 sampai 5

)

* 1. **Preprocessing**

*Pre-processing* dilakukan untuk membersihkan data komentar, tahap ini membuat data yang tidak baik atau belum siap diolah menjadi data yang siap diubah. tahap ini terdiri dari beberapa tahapan yaitu *case folding*, *data cleaning*, tokenisasi, *stopword*, *stemming* dan normalisasi kata tidak baku. Diagram proses dari *pre-processing* dapat dilihat pada Gambar 3.2.



Gambar 3.2 Proses *Preprocessing*