

**ANALISIS SENTIMEN MENGGUNAKAN A LITE BERT PADA OPINI
MASYARAKAT DI MEDIA SOSIAL X TERKAIT KAMPANYE HINGGA
PELANTIKAN PRESIDEN 2024 MENGGUNAKAN SUPPORT
VECTOR MACHINE**

PROPOSAL JUDUL

Diajukan kepada Program Studi Informatika
Untuk Menyusun Skripsi



Oleh :

RAFIF ILAFI WAHYU GUNAWAN
NPM. 21081010093

**PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN” JAWA TIMUR
2024**

LEMBAR PENGESAHAN JUDUL

**ANALISIS SENTIMEN MENGGUNAKAN A LITE BERT PADA OPINI
MASYARAKAT DI MEDIA SOSIAL X TERKAIT KAMPANYE HINGGA
PELANTIKAN PRESIDEN 2024 MENGGUNAKAN SUPPORT
VECTOR MACHINE**

Oleh :

Rafif Ilafi Wahyu Gunawan
NPM. 21081010093

Telah diterima pada tanggal

.....bulan.....Tahun

KATA PENGANTAR

Puji syukur Alhamdulillah kehadiran Allah SWT yang telah memberikan rahmat dan hidayah-Nya kepada penulis sehingga dapat menyusun dan menyelesaikan proposal terkait pengajuan judul yang akan digunakan untuk skripsi dengan judul “Analisis Sentimen Menggunakan A Lite Bert pada Opini Masyarakat di Media Sosial X Terkait Kampanye hingga Pelantikan Presiden 2024 Menggunakan Support Vector Machine” dengan lancar. Penyusunan proposal pengajuan judul ini dibuat sebagai syarat untuk bisa memilih mata kuliah pra-skripsi (SEMPRO) sebagai tahapan pertama di Program Studi Informatika Fakultas Ilmu Komputer, Universitas Pembangunan Nasional “Veteran” Jawa Timur.

Penulis juga berterima kasih kepada Ibu Eva Yulia Puspaningrum, S.Kom., M.Kom selaku dosen mata kuliah data mining yang telah menyediakan waktu, tenaga, dan pikiran untuk mengarahkan serta memotivasi penulis dalam penyusunan proposal pengajuan judul ini serta Dr. Basuki Rahmat, S.Si. MT selaku dosen mata kuliah riset informatika yang telah membantu penulis melalui saran dan kritik yang diberikan, terima kasih atas bimbingannya yang telah diberikan. Tidak lupa penulis juga mengucapkan terima kasih kepada pihak – pihak terkait atas segala bantuan, dukungan serta bimbingan yang telah diberikan selama proses pembuatan proposal skripsi ini, yaitu :

1. Prof. Dr. Ir. Novirina Hendrasarie, M.T. selaku Dekan Fakultas Ilmu Komputer Universitas Pembangunan Nasional “Veteran” Jawa Timur
2. Fetty Tri Anggrainy, S.Kom. selaku Ketua Jurusan Program Studi Agribisnis Fakultas Pertanian Universitas Pembangunan Nasional “Veteran” Jawa Timur

3. Fetty Tri Anggrainy, S.Kom., M.Kom. selaku Koordinator Program Studi Informatika Fakultas Ilmu Komputer Universitas Pembangunan Nasional “Veteran” Jawa Timur
4. Kedua Orang Tua penulis tercinta yang senantiasa memberikan doa, semangat, dan motivasi bagi penulis untuk menyelesaikan penyusunan proposal judul skripsi
5. Keluarga dan teman-teman tersayang yang senantiasa memberikan dukungan selama pembuatan proposal skripsi
6. Seluruh pihak yang terlibat dalam pelaksanaan dan pembuatan proposal judul skripsi

Penulis menyadari jika masih banyak kekurangan dalam penyusunan proposal judul skripsi ini. Oleh karena itu penulis menerima segala kritik dan saran yang membangun untuk perbaikan di masa depan. Akhir kata, dengan selesainya penulisan proposal judul skripsi ini memberikan manfaat bagi penulis khususnya dan umunya dapat membantu berbagai pihak yang membutuhkan informasi yang berkaitan dengan tulisan ini.

Surabaya, 23 Desember 2024

Rafif Ilafi Wahyu Gunawan

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

I. PENDAHULUAN

1.1 Latar Belakang

Media sosial X merupakan aplikasi atau platform yang memiliki pengguna sekitar 24 juta di Indonesia, media sosial X menjadi salah satu alat yang praktis dan efektif dalam menyebar luaskan informasi secara publik, sehingga platform ini sering digunakan oleh masyarakat Indonesia dalam beropini secara terbuka yaitu salah satunya politik (**Kutipan aplikasi X**). Politik salah satu topik yang sering digunakan oleh masyarakat Indonesia dalam menyuarakan opini, Seringkali masyarakat memiliki kebebasan berpendapat dan berekspresi di media sosial yang tercantum pada UU No 19 Tahun 2016 dan UU ITE pasal 27 ayat (3) menjadi dasar penting dalam membangun diskusi terbuka di platform ini (**Kutipan Opini**). Melihat peran penting media sosial X dalam mendukung kebebasan berekspresi, platform ini dapat menjadi sarana yang efektif untuk menyampaikan informasi yang berkaitan dengan proses demokrasi, termasuk pemilihan presiden. Di tahun 2024 momentum pesta demokrasi yang akan menentukan masa depan bangsa, media sosial X menjadi salah satu ruang strategis bagi berbagai pihak untuk menyampaikan visi, gagasan, dan program yang diharapkan dapat memberikan dampak positif bagi masyarakat. Dan memberikan ruang bagi masyarakat untuk berpartisipasi aktif dalam proses demokrasi baik menyampaikan opini maupun gagasan, serta presiden dan wakil presiden mempunyai hak melaksanakan kampanye yang tercantum pada UU No 7 Tahun 2017 pasal 299 dalam mendapatkan suara dan hati masyarakat(**Kutipan Kampanye**). Dari hasil kampanye yang efektif membantu masyarakat menentukan pilihan dalam pemilihan

presiden, diikuti dengan pemungutan suara, hingga pelantikan presiden dan wakil presiden terpilih dalam sidang MPR sebagai awal masa pemerintahan baru **(Kutipan Pemilihan Presiden)**.

Dari fenomena pemilihan presiden, opini menjadi sumber informasi yang sangat berharga untuk dilakukan analisis menggunakan teknologi. *Data mining* yang menerapkan teknologi pemrosesan bahasa alami (*natural language processing*) dan pembelajaran mesin (*machine learning*) dalam studi analisis sentimen terdiri dari tiga kelas diantaranya, sentimen positif yang merupakan dukunagan terhadap para kandidat. Sentimen negatif merupakan bentuk kritik atau kekhawatiran dan ketidakpuasan masyarakat terhadap para kandidat, serta isu – isu sensitif yang dapat memengaruhi opini publik dan memicu polarisasi. Sentimen netral merupakan kecenderungan masyarakat terhadap opini yang tidak ke arah positif atau negatif terhadap kandidat, dari data sentimen masyarakat terhadap para kandidat diidentifikasi untuk memberikan wawasan penting yang dapat dimanfaatkan oleh tim kandidat dalam pemilihan presiden selanjutnya dalam menyusun strategi komunikasi yang lebih efektif selama proses pemilihan berlangsung **(Kutipan Analisis Sentimen)**.

Terdapat beberapa algoritma yang digunakan untuk melakukan analisis sentimen dengan berbagai studi terkait isu – isu yang terjadi di masyarakat pada sosial media, salah satunya penelitian yang dilakukan oleh Cahyani Putri & Abdillah Lufty pada tahun 2024 berjudul “Perbandingan Performa Algoritma *Naïve Bayes*, SVM dan *Random Forest*: Studi Kasus Analisis Sentimen Pengguna Sosia Media X”**(Kutipan Perbadingan Algoritma)**. Pada penelitian ini bertujuan membandingkan perfoma dari ketiga algoritma yaitu *Naïve Bayes*, SVM, *Random*

Forest, dari ketiga algoritma tersebut dalam mengevaluasi terkait opini pengguna media sosial terhadap Ibu Kota Nusantara (IKN) melalui platform X. Dengan jumlah data tweet sebanyak 10.000 menggunakan teknik crawling data menunjukkan hasil performa algoritma, bahwa SVM memiliki akurasi performa yang paling tertinggi dari algoritma *Naïve Bayes* dan *Radom Forest*, dengan akurasi, presisi, recall dan f-1 score sebesar 87% sedangkan algoritma *Naïve Bayes* 73% dengan selisih kinerja sebesar 14%, sedangkan untuk algoritma *Radom Forest* 85% dengan selisih kinerja sebesar 2%. Algoritma SVM menunjukkan performa dengan bagus dalam memahami opini masyarakat di media sosial terhadap IKN.

Dalam konteks penelitian sebelumnya SVM terpilih menjadi akurasi yang paling tertinggi dari dua algoritma *Naïve Bayes* dan *Random Forest* dengan menggunakan TF-IDF sebagai ekstraksi fitur yang mengubah dari data teks menjadi representasi numerik, yang memungkinkan komputer untuk memahami hubungan antar entitas dalam pengelompokan data dan juga memungkinkan data visualisasi dalam bentuk grafik dan confusion matrix. Oleh karena itu penelitian ini ingin mengimplementasikan sebuah metode ekstraksi fitur yaitu A Lite BERT sebagai model dalam analisis sentimen. A Lite BERT (ALBERT) memiliki tugas yang sama seperti TF-IDF yaitu sebagai pemrosesan bahasa alami (*Natural Language Processing*) dan memahami antar kata maupun kalimat, salah satunya penelitian yang dilakukan oleh Lee et al., pada tahun 2021 berjudul “KoreALBERT: Pretraining a Lite BERT Model for Korean Language Understanding” (**Kutipan A Lite BERT**). Pada penelitian ini sebagai *Masked Language Modelling* (MLM) yang membantu model memahami konteks antar kata dengan lebih mendalam, A Lite BERT juga berguna dalam *Sentence Order Prediction* (SOP) yang dirancang untuk

mempelajari antar kalimat, dan A Lite BERT berguna dalam *Word Order Prediction* (WOP) yang berfokus pada prediksi urutan kata dalam kalimat dalam menangani kompleksitas bahasa aglutinatif seperti Korea.

Dari latar belakang yang telah dijelaskan sebelumnya. Penelitian ini akan mengimplementasikan metode A Lite BERT (ALBERT) menggunakan algoritma SVM yang berjudul “Analisis Sentimen Menggunakan A Lite Bert pada Opini Masyarakat di Media Sosial X Terkait Kampanye hingga Pelantikan Presiden 2024 Menggunakan Support Vector Machine” Tujuan penelitian ini adalah untuk mengetahui seberapa pengaruhnya menggunakan metode A Lite BERT digunakan sebagai ekstraksi fitur dalam analisis sentimen pada data teks berupa opini masyarakat di media sosial X pada fenomena kampanye hingga pelantikan presiden 2024. Dengan demikian, penelitian ini akan mengevaluasi hasil kinerja yang digunakan dari sentimen masyarakat terhadap fenomena tersebut dengan menggunakan algoritma *Support Vector Machine* (SVM).

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan sebelumnya, penulis merumuskan pertanyaan penelitian sebagai berikut: “Bagaimana performa ALBERT dalam melakukan analisis sentimen opini masyarakat di media sosial X terkait kampanye hingga pelantikan presiden 2024 menggunakan algoritma *support vector machine*?” penjabaran lebih rinci mengenai permasalahan tersebut sebagai berikut :

1. Bagaimana alur proses ekstraksi fitur menggunakan model ALBERT dalam menganalisis sentimen opini masyarakat di media sosial X terkait kampanye hingga pelantikan Presiden 2024 dengan algoritma Support Vector Machine?

2. Bagaimana performa model ALBERT dalam melakukan analisis sentimen terhadap opini masyarakat di media sosial X terkait kampanye hingga pelantikan Presiden 2024 menggunakan algoritma Support Vector Machine?
3. Sejauh mana model ALBERT sebagai metode ekstraksi fitur dapat mendukung akurasi klasifikasi sentimen menggunakan algoritma Support Vector Machine pada opini masyarakat di media sosial X terkait kampanye hingga pelantikan Presiden 2024?

1.3 Batasan Masalah

Dalam penelitian ini, batasan masalah ditetapkan untuk menjaga fokus analisis dan memastikan penelitian dilakukan secara sistematis sesuai dengan tujuan yang telah dirumuskan. Batasan ini dirancang untuk memperjelas ruang lingkup penelitian dan membatasi variabel yang tidak relevan agar hasil penelitian lebih terarah dan valid. Beberapa batasan masalah sebagai berikut:

1. Penelitian ini menggunakan data tweet berbahasa Indonesia yang didapatkan di media sosial twitter dengan topik yaitu “Kampanye Presiden 2024 Sampai Pelantikan Presiden 2024”.
2. Dataset untuk topik “Kampanye Presiden 2024 Sampai Pelantikan Presiden 2024” dengan rentang 28 November 2023 sampai 20 Oktober 2024
3. Penelitian Menggunakan 10.000 data yang akan dibagi 80% data latih 20% data uji
4. Penelitian menggunakan metode ekstraksi fitur ALBERT .
5. Bahasa yang digunakan dalam implementasi penelitian ini menggunakan pemrograman Python.

6. Analisis sentimen terhadap opini masyarakat di media sosial x terkait kampanye hingga pelantikan presiden 2024 dengan tiga kategori 0 negatif, 1 positif, 2 netral.

1.4 Tujuan Penelitian

Berdasarkan latar belakang dan rumusan masalah diatas, maka dapat dirumuskan tujuan penelitian ini adalah untuk mengetahui analisis sentimen terhadap opini masyarakat di media sosial X terkait kampanye hingga pemilihan presiden 2024 menggunakan ekstraksi fitur ALBERT saat diimplementasikan menggunakan Support Vector Machine. Berikut penjelasan detail penjelasan dari penelitian ini:

1. Mendeskripsikan alur proses ekstraksi fitur menggunakan model ALBERT, mulai dari preprocessing data teks hingga pengintegrasian fitur dengan algoritma Support Vector Machine untuk klasifikasi sentimen.
2. Mengevaluasi performa model ALBERT dalam melakukan analisis sentimen, berdasarkan metrik akurasi, precision, recall, dan F1-score pada opini masyarakat di media sosial X.
3. Menganalisis kontribusi model ALBERT sebagai metode ekstraksi fitur dalam mendukung akurasi klasifikasi sentimen menggunakan algoritma Support Vector Machine.

1.5 Manfaat Penelitian

Penelitian ini nantinya dapat bermanfaat bagi peneliti, pembaca, perguruan tinggi, masyarakat umu, dan partai polotik. Berikut beberapa manfaat yang diharapkan yaitu :

1. Bagi peneliti, sebagai sarana pengimplementasi pengetahuan serta pelatihan dalam menerapkan ilmu yang telah dipelajari selama perkuliahan.
2. Bagi pembaca, sebagai sarana sumber informasi dan inspirasi untuk penelitian selanjutnya untuk memperluas dalam pemahaman terkait isu-isu politik terhadap pola analisis masyarakat di media sosial twitter terkait kampanye dan pemilihan presiden di Indonesia .
3. Bagi perguruan tinggi, sebagai sumber referensi, informasi dan tambahan pengetahuan untuk penelitian lebih lanjut mengenai identifikasi isu-isu utama yang menjadi perhatian publik, membantu pembuatan kebijakan yang lebih responsif terhadap kebutuhan masyarakat.
4. Bagi partai politik, sebagai tambahan informasi serta pemahaman dalam memahami sentimen dan respon masyarakat, sehingga partai politik dapat menyusun strategi komunikasi yang lebih tepat sasaran berdasarkan data yang sudah dianalisis di media sosial.
5. Bagi Masyarakat Umum, sebagai gambaran tentang bagaimana respons masyarakat terhadap isu-isu politik dapat dianalisis secara objektif, sehingga meningkatkan kesadaran terhadap dinamika opini publik.

II. TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian terdahulu merupakan salah satu acuan peneliti dalam melakukan penelitian sehingga penulis dapat memperkaya teori untuk mengkaji penelitian yang dilakukan. Berdasarkan penelitian terdahulu, penulis mengangkat beberapa penelitian sebagai referensi dalam memperkaya bahan kajian penulis serta untuk mencari bahan perbandingan yang selanjutnya menemukan inspirasi baru untuk penelitian. Dalam mendukung permasalahan terhadap bahasan tersebut, peneliti berusaha mencari berbagai literatur dan penelitian terdahulu yang masih relevan terhadap masalah yang menjadi obyek penelitian. Berikut beberapa mapping yang dilakukan dan ditemukan beberapa penelitian yang dijadikan acuan oleh penulis:

1. Pada penelitian yang dilakukan oleh () dengan judul penelitian "Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture". Penelitian ini membahas perbandingan kinerja pada 4 *machine learning*, yaitu *Support Vector Machine* (SVM), *Naïve Bayes* (NB), *Random Ferns* (RFerns), dan *K-Nearest Neighbor* (kNN) dalam mendeteksi parameter fisiko-kimia air yang berpotensi menyebabkan penyakit pada akuakultur ikan yang diklasifikasikan menjadi enam kelompok berdasarkan keberadaan bakteri. Data dibagi menjadi 75% untuk pelatihan dan 25% untuk pengujian, dengan evaluasi menggunakan metrik akurasi, kappa, sensitivitas, spesifisitas. Mendapatkan bahwa *machine learning* SVM dan Rferns adalah metode yang paling efektif untuk mendeteksi parameter penyebab penyakit dalam akuakultur dengan mencapai akurasi, recall, presisi, *f1-score* mencapai skor 100%

2. Pada penelitian yang dilakukan oleh () dengan judul penelitian “Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset”. Penelitian ini membahas perbandingan kinerja pada beberapa algoritma *Naive Bayes* dan SVM yang digunakan klasifikasi maupun analisis sentimen, dengan data yang digunakan sebesar 10.000 ulasan maskapai penerbangan dari twitter yang dibagi menjadi 67% data latih dan 33% data uji untuk pengujian sentimen dengan label (positif, negatif, dan netral). Dalam penelitian ini menggunakan *confusion matrix* untuk mengevaluasi akurasi, presisi, recall, *f1-score*. Pada Algoritma SVM memiliki Akurasi 82,48%, presisi 90,33%, recall 81,79%, F1-score 85,85%. SVM lebih unggul 5,92% akurasi.
3. Pada penelitian yang dilakukan oleh () dengan judul penelitian “Performance Analysis of Transformer Based Models (BERT, ALBERT and RoBERTa) in Fake News Detection”. Penelitian ini membahas perbandingan kinerja pada beberapa model berbasis transformer seperti BERT, ALBERT, dan RoBERTa dalam mendeteksi berita palsu di bahasa Indonesia dengan dataset yang terdiri dari *TurnHoax*-dataset: 433 berita valid dan 683 berita palsu, dataset dari mendeley: 372 berita valid dan 228 berita palsu, dan *Hoax-NewsClassification* dataset: 250 berita valid dan 250 berita palsu. Menunjukkan keunggulan dalam tugas NLP dalam mendeteksi berita palsu di Indoensia model ALBERT memiliki kinerja terbaik dengan akurasi 87.6%, presisi 86.9%, dan *f1-score* 86.9%.
4. Pada penelitian yang dilakukan oleh () dengan judul penelitian “Implementation of a Machine Learning Algorithm for Sentiment Analysis

of Indonesia's 2019 Presidential Election". Penelitian ini membahas perbandingan kinerja pada dari dua algoritma yaitu *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM) dengan teknik tokenisasi yang meliputi *Alphabetic Tokenizer*, *Character N-gram*, *Unigram*, *Bigram*, *Trigram*, *N-gram*, dan *Word Tokenizer* pada analisis sentimen. Dengan menggunakan 10.000 dari tweet dalam bahasa Indonesia terdiri dari 5.000 tweet dengan kata kunci "jokowi" dan 5000 tweet dengan kata kunci "prabowo". Algoritma SVM dengan tokenisasi menggunakan alfabet menunjukkan hasil dengan akurasi, presisi, recall tertinggi sebesar 79.02%.

5. Pada penelitian yang dilakukan oleh () dengan judul penelitian "A Novel Named Entity Recognition Scheme for Steel E-Commerce Platforms Using a Lite BERT." Penelitian ini membahas pengenalan entitas bernama (NER) berbasis *deep learning* menggunakan ALBERT, dengan membandingkan model BERT standar. Dataset yang digunakan berisi terdiri dari 1322 data historis pembelian bahasa yang telah dibalebli oleh pakar baja menggunakan skema anotasi yang berbeda. Penelitian ini merancang ALBERT sebagai ekstraksi fitur juga dengan menggunakan teknik *embedding* dalam merepresentasikan tugas NER untuk memahami entitas dalam konteks spesifiknya. Representasi ini membantu menangkap fitur – fitur teks yang relevan untuk tugas NER. Dari model ALBERT dengan fungsi loss baru menunjukkan hasil terbaik dibandingkan metode lain termasuk BERT-CRF dan Word2Vec-BiLSTM-CRF. Dengan skema IO, precision mencapai 90%, recall 91.3%, dan F1-score 0.91. Model ini juga mengurangi waktu pelatihan dibandingkan model BERT, membuatnya lebih efisien.

2.2 X

Platform media sosial *Twitter* yang kini dikenal dengan nama X, merupakan sebuah media sosial yang sangat populer dan banyak digunakan kalangan masyarakat. Termasuk Indonesia dengan jumlah total pengguna sebesar 14,58 juta pada tahun 2022. X merupakan tempat semua orang bisa bebas memberikan cuitan atau tweet secara bebas baik itu positif maupun cuitan negatif tanpa ragu. Hal ini didasari oleh X itu sendiri dalam menyediakan opsional pada sebuah *settingan* pada akun pengguna dalam memprivat identitas. Kepemilikan akun privat tersebut menjadi dorongan banyak orang dalam bebas melakukan cuitan tanpa memikirkan dampak bagi sekitarnya (**Kutipan X**).

Aplikasi X menjadi pilihan konsumsi masyarakat dalam bermedia sosial, hal itu karena platform ini sangat efektif untuk menjadi tempat menjalin sebuah interaksi atau hubungan dengan orang lain sebagai teman di berbagai belahan dunia. Karena platform ini memberikan kemudahan bagi para pengguna untuk bisa saling mengirim pesan tanpa batas, mengirim gambar maupun video. Dan X menjadi platform media sosial di Indonesia tercepat dalam mendapatkan berbagai informasi terkini. Hal ini didasari karena X memiliki fitur berupa “tren”, yang menjadi dasar informasi masyarakat dalam pemenuhan berita terbaru (**Kutipan X2**).

2.3 Pemilihan Presiden

Pemilihan Presiden atau sering dikenal dengan pilpres, merupakan sebuah pemilihan yang digunakan dalam mewujudkan kekuasaan rakyat yang sudah diatur dalam Undang – Undang Dasar 1945. Oleh karena itu dalam pemilihan umum diterapkan asas *luber jurdil* (langsung, umum, bebas, rahasia, jujur, dan adil) merupakan perintah konstitusi yang harus dilaksanakan. Rakyat merupakan

pemegang kedaulatan tertinggi, bertindak dalam memberikan kekuasaanya kepada lembaga seperti Presiden, DPR, DPD, dan DPRD melalui pemilihan Umum. Pemilu juga bisa menjadi tempat dimana individu maupun golongan berinteraksi untuk mengadakan kesepakatan sosial dari parpol terhadap pemegang suara secara sah yakni masyarakat sipil (**Kutipan Pemilihan Presiden**). Dalam pemilihan presiden meliputi diantaranya:

2.3.1 Kampanye Pemilu

Kampanye pemilu adalah aktivitas terorganisir untuk menyampaikan visi, misi, dan program kandidat guna memengaruhi pilihan masyarakat. Berdasarkan Undang-Undang Nomor 42 Tahun 2008, kampanye bertujuan mendukung proses demokrasi yang sehat melalui metode seperti debat, media massa, dan pertemuan. Dalam Islam, kampanye dianggap sebagai hak politik individu, tetapi harus dijalankan secara etis, tanpa memfitnah atau merendahkan pihak lain, melainkan fokus pada penyampaian gagasan yang konstruktif untuk kepentingan masyarakat (**Kutipan Kampanye Pemilu**).

2.3.2 Pencoblosan Pemilu

Pencoblosan dalam pemilu adalah bentuk partisipasi politik aktif yang memungkinkan warga negara memilih pemimpin berdasarkan visi, misi, dan program kerja kandidat. Meskipun prosesnya singkat, dampaknya berlangsung hingga lima tahun ke depan, sehingga keputusan harus dibuat dengan cermat dan rasional. Dengan mencoblos, pemilih tidak hanya berkontribusi dalam demokrasi tetapi juga menentukan masa depan pemerintahan yang berkualitas (**Kutipan Pencoblosan Pemilu**).

2.3.3 Pemungutan Suara

Pemungutan suara di Indonesia secara tradisional menggunakan kertas suara dengan proses manual yang memakan waktu, rawan kesalahan, dan mahal. E-voting diusulkan sebagai alternatif untuk meningkatkan efisiensi dan akurasi dengan menggunakan perangkat elektronik. Namun, penerapannya terkendala kesiapan infrastruktur, sumber daya manusia, dan penerimaan masyarakat terhadap teknologi baru. Transisi bertahap dapat dilakukan, dimulai dari wilayah yang lebih siap secara teknologi (**Kutipan Pemungutan Suara**).

2.3.4 Pelantikan Presiden

Pelantikan Presiden Indonesia diatur dalam Pasal 9 UUD 1945, di mana presiden mengucapkan sumpah jabatan di hadapan MPR atau DPR. Jika keduanya berhalangan, sumpah dilakukan di hadapan pimpinan MPR dan disaksikan pimpinan MA. Mekanisme ini lebih bersifat seremonial, dengan MPR sebagai saksi. Isi sumpah menegaskan janji untuk mematuhi UUD, menjalankan tugas secara adil, dan berbakti kepada bangsa, dengan tanggung jawab moral utama kepada rakyat sebagai pemberi mandat (**Kutipan Pelantikan Presiden**).

2.4 Analisis Sentimen

Analisis sentimen adalah proses untuk mengelompokkan dari sebuah opini menjadi kategori atau label dengan nilai positif, negatif, atau netral. Meskipun sering diterapkan pada opini publik di media sosial, analisis ini juga dapat dilakukan pada berbagai sumber data yang bisa digunakan, seperti ulasan produk, survei pelanggan, atau komentar di platform media sosial lainnya. Algoritma yang digunakan mampu mengidentifikasi pola dan emosi dalam data secara efektif. Hasil analisis ini memberikan wawasan yang berguna bagi entita seperti perusahaan atau





























pemerintah untuk memahami persepsi masyarakat, dengan tingkat akurasi yang dipengaruhi oleh kualitas data pelatihan dan parameter algoritma yang diterapkan **(Kutipan Analisis Sentimen)** .

2.5 FlowChart

Flowchart memiliki peran yang sangat penting dalam mendukung sistem alur kerja karena mampu memberikan gambaran visual yang terstruktur mengenai langkah-langkah yang harus dijalankan dalam sebuah proses. Dengan adanya flowchart, setiap tahapan dalam alur kerja dapat diikuti dengan lebih mudah, sehingga pengguna dapat memahami setiap detail proses tanpa kesulitan. Selain itu, flowchart membantu dalam mengidentifikasi potensi masalah atau kendala yang mungkin muncul, sehingga memungkinkan langkah pencegahan atau perbaikan dilakukan dengan lebih cepat dan efisien.

Flowchart juga berfungsi sebagai alat untuk mengurangi kebingungan di antara para pelaku proses dengan memberikan panduan yang jelas mengenai urutan dan tanggung jawab pada setiap tahap. Dalam konteks kerja tim, flowchart menjadi sarana komunikasi yang efektif, memastikan bahwa semua anggota tim memiliki pemahaman yang sama terkait tugas dan alur kerja yang sedang dilakukan. Hal ini juga mendukung pengambilan keputusan yang lebih terarah karena prosedur yang digambarkan secara grafis dapat dijadikan acuan yang mudah dimengerti. Oleh karena itu, flowchart menjadi alat yang esensial dalam proses perancangan, analisis,

hingga optimalisasi sistem alur kerja agar dapat berjalan lebih efisien dan terorganisir (**Kutipan Flowchart**).

| Gambar | Fungsi | Gambar | Fungsi |
|---|---------------------------------|---|---------------------------|
|  | Proses |  | Card |
|  | Proses pilihan |  | Punched tape |
|  | Keputusan |  | Summing Junction |
|  | Input Data dan Output Informasi |  | Or |
|  | Predefine Proses |  | Collate |
|  | Internal Storage |  | Sort |
|  | Dokumen |  | Extract |
|  | MultiDokumen |  | Merge |
|  | Terminator (mulai dan Akhir) |  | Storage Data |
|  | Preparasi |  | Delay |
|  | Manual Input |  | Sequential Access Storage |
|  | Manual Operasi |  | Magnetic Disk |
|  | Penghubung |  | Direct Access Storage |
|  | Off Page Penghubung |  | Display |

2.6 Data Mining

Data mining adalah metode terstruktur yang digunakan untuk mengeksplorasi dan menemukan pola serta hubungan dalam kumpulan data yang besar. Proses ini dirancang untuk membuat model prediktif dari data historis guna mendukung pengambilan keputusan di masa depan. Data mining melibatkan penggunaan pendekatan seperti statistik, algoritma pembelajaran mesin (*Machine Learning*), dan teknologi kecerdasan buatan (*Artificial Intelligence*) untuk mengolah data mentah menjadi data yang siap digunakan untuk kebutuhan analisis mendalam

untuk mendapatkan informasi yang bermakna dan relevansi (**Kutipan Data Mining**).

2.7 Pelabelan Data (*Labelling*)

Pelabelan data (*Labelling*) merupakan proses yang menentukan jenis label atau kelas yang akan di gunakan sesuai dengan kebutuhan. Misalnya pada analisis sentimen yaitu dengan kelas berupa positif, negatif, atau netral. Dari label tersebut untuk setiap kelasnya akan dilakukan pengkategorian berdasarkan jenis dari data tersebut jika berbentuk teks maka akan dianalisis dengan mencocokkan kata-kata di dalamnya dengan kamus sentimen yang mengandung nilai polaritas tertentu. Nilai-nilai ini dijumlahkan untuk menentukan kategori sentimen, yang kemudian digunakan sebagai label untuk teks tersebut.

Metode *Lexicon-Based* menilai sentimen teks dengan memanfaatkan kamus yang memuat daftar kata beserta nilai polaritasnya. Proses dimulai dengan mengumpulkan data dari media sosial, diikuti dengan pemrosesan data untuk membersihkan elemen-elemen yang tidak relevan. Setelah itu, kata-kata kunci diekstrak dan dicocokkan dengan entri kamus untuk menentukan polaritasnya. Hasilnya digunakan untuk mengklasifikasikan sentimen teks menjadi positif, negatif, atau netral. Berikut rumus *lexicon based* (**Kutipan Pelabelan Data**).

$$\text{Nilai Sentimen} = \begin{cases} 1 & \text{jika } \sum \text{Polarittas} > 0 \\ 0 & \text{jika } \sum \text{Polarittas} = 0 \\ -1 & \text{jika } \sum \text{Polarittas} < 0 \end{cases}$$

Penjelasan pada rumus tersebut sebagai berikut

- $\sum \text{Polaritas}$: Menjumlahkan nilai polaritas dari setiap kata dalam teks. Polaritas dapat berupa nilai positif, negatif, atau nol yang diperoleh dari kamus sentimen.

- $\sum \text{Polarittas} > 0$: jika total polaritas lebih besar dari 0 maka label positif.
- $\sum \text{Polarittas} = 0$: jika total polaritas sama dengan 0 maka label netral.
- $\sum \text{Polarittas} < 0$: jika total polaritas lebih kecil dari 0 maka label negatif.

2.8 Pemrosesan Data (*Preprocessing*)

Pemrosesan data (*Preprocessing*) merupakan rangkaian langkah yang dilakukan untuk mengolah data mentah agar lebih terorganisir, terstruktur, dan siap digunakan dalam proses analisis atau pembuatan model. Dalam konteks pengolahan teks, *preprocessing* memiliki tujuan utama untuk menyeleksi informasi yang dianggap relevan sekaligus mengeliminasi elemen-elemen yang tidak diperlukan. Dengan menyaring elemen-elemen tersebut, data yang dihasilkan menjadi lebih sederhana dan fokus pada bagian-bagian penting yang mendukung analisis lebih mendalam. Proses ini tidak hanya membantu menyederhanakan data tetapi juga berkontribusi secara signifikan terhadap peningkatan efisiensi serta akurasi algoritma yang akan digunakan dalam proses klasifikasi atau analisis lebih lanjut (**Kutipan Preporcessing**). Berikut beberapa tahapan pada *preprocessing* diantaranya:

2.8.1 *Cleansing*

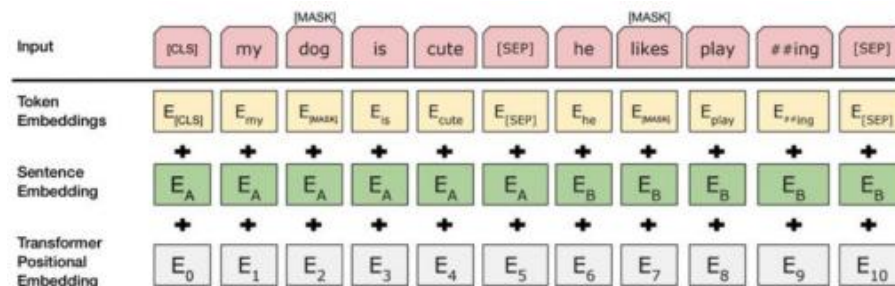
Cleansing Data cleansing adalah proses untuk mengidentifikasi dan memperbaiki kesalahan, inkonsistensi, atau data yang tidak relevan, dengan tujuan meningkatkan kualitas data. Tahapannya meliputi analisis data, transformasi, dan penggantian data mentah dengan versi yang telah dibersihkan. Proses ini memastikan data siap digunakan untuk analisis atau pemodelan yang akurat (**Kutipan Cleansing**).

2.8.2 Case Folding

Case folding Case folding adalah proses dalam pengolahan teks yang mengubah semua huruf dalam dokumen menjadi huruf kecil dan menghapus karakter non-alfabet, seperti angka, tanda baca, atau simbol lainnya. Langkah ini bertujuan untuk menyamakan format teks sehingga lebih mudah dianalisis oleh algoritma pemrosesan teks. Dengan cara ini, perbedaan huruf besar dan kecil tidak memengaruhi hasil pengolahan data (**Kutipan Case Folding**).

2.8.3 Tokenizing

Tokenizing merupakan langkah dalam pengolahan teks yang bertujuan memecah sebuah teks menjadi bagian-bagian lebih kecil, seperti kata-kata atau kelompok kata. Proses ini membantu menyusun teks yang lebih terstruktur sehingga dapat dianalisis dengan lebih efisien. Langkah ini penting untuk menyiapkan data teks sebelum dimasukkan ke dalam model pembelajaran mesin atau kecerdasan buatan lainnya (**Kutipan Tokenizing**).



Gambar () pada proses tokenisasi dan embedding dalam model BERT, yang terdiri dari token embedding, sentence embedding, dan positional embedding. Dalam model ALBERT, proses ini serupa, namun dengan modifikasi yang signifikan untuk meningkatkan efisiensi. ALBERT menggunakan factorized embedding parameterization, yang memisahkan ukuran embedding token

dari hidden layer transformer, mengurangi parameter embedding tanpa mengorbankan kinerja. Selain itu, ALBERT mengganti tugas NSP dengan SOP untuk meningkatkan pemahaman model terhadap hubungan antar kalimat."

2.8.4 *Encoding*

Encoding adalah langkah awal dalam proses komunikasi di mana pengirim pesan, seperti produser media, merancang makna tertentu dalam pesan mereka. Ini dilakukan berdasarkan konteks sosial, politik, dan budaya mereka sendiri. Dalam pandangan Stuart Hall, pesan ini tidak hanya berupa informasi biasa tetapi juga mencerminkan nilai-nilai ideologis yang mungkin sengaja atau tidak sengaja disematkan oleh pengirimnya (**Kutipan Encoding**).

Berikut rumus dalam *encoding*:

$$IG(S,A) = H(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} H(S_v)$$

Penjelasan pada rumus tersebut sebagai berikut:

- $H(S)$: Entropi dataset sebelum pembagian.
- $|S_v|$: Jumlah sampel di subset v.
- $|S|$: Total jumlah sampel.
- $H(S_v)$: Entropi subset setelah pembagian.

2.9 Membagi Data (*Split Data*)

Split Data adalah proses pemisahan data bertujuan untuk membagi dataset menjadi dua kelompok: data pelatihan dan data pengujian. Umumnya, 70% atau 80% dataset digunakan untuk melatih model, sedangkan 30% atau 20% sisanya digunakan untuk menguji kinerja model. Pendekatan ini membantu menilai

kemampuan model dalam memproses data baru yang belum pernah digunakan sebelumnya.

K-Fold Cross Validation adalah pendekatan statistik untuk menilai kinerja model dalam machine learning dengan membagi dataset menjadi beberapa bagian (fold). Setiap bagian bergantian berfungsi sebagai data pengujian, sementara bagian lainnya digunakan untuk melatih model. Metode ini efektif dalam mengurangi bias, mengatasi overfitting, dan menghasilkan evaluasi yang lebih stabil. Sebagai contoh, dengan 10-Fold Cross Validation, dataset dibagi menjadi sepuluh bagian yang masing-masing bergantian menjadi data pengujian dalam sepuluh iterasi. Hasilnya dihitung berdasarkan rata-rata akurasi dari semua pengujian (**Kutipan K-Fold Cross Validation**). Rumus *K-Fold Cross Validation* sebagai berikut

$$\text{Akurasi rata – rata} = \frac{1}{k} \sum_{i=1}^k \text{Akurasi}_i$$

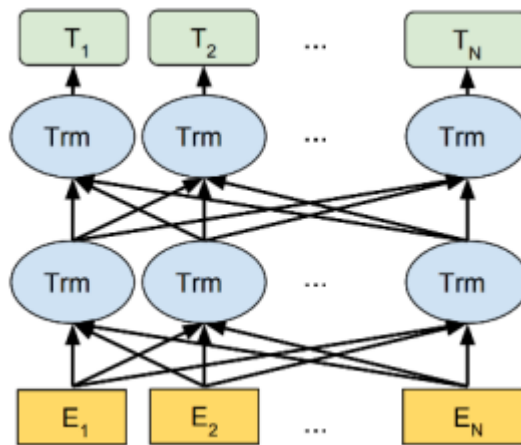
Penjelas pada rumus tersebut sebagai berikut:

- k : Jumlah *fold* (bagian dataset yang digunakan dalam validasi)
- Akurasi_i : Akurasi yang diperoleh pada *fold* ke- i
- $\frac{1}{k}$: Faktor rata – rata yang membagi jumlah akurasi berdasarkan jumlah

fold

2.10 Model *A Lite BERT* (ALBERT)

A Lite BERT merupakan varian ringan dari *BERT* yang dirancang untuk meningkatkan efisiensi dan performa dengan mengurangi ukuran parameter melalui teknik faktorisasi embedding dan berbagi parameter antar-lapisan. Dengan mengganti metode prediksi antar-kalimat menjadi *SOP*, *ALBERT* berhasil meningkatkan koherensi antar-kalimat, menghasilkan performa yang lebih tinggi dengan kebutuhan sumber daya yang lebih rendah (**Kutipan ALBERT**).



Pada gambar () menunjukkan arsitekti dasar model BERT (*Bidirectional Encoder Representations from Transformers*), yang juga menjadi inti dari model *A Lite BERT* (ALBERT). ALBERT mempertahankan struktur dasar transformer, namun memperkenalkan beberapa optimisasi untuk meningkatkan efisiensi. Di antaranya adalah parameter *sharing* antar lapisan *transformer*, *factorized embedding parameterization* untuk mengurangi ukuran parameter *embedding*, dan *sentence-order prediction loss* untuk meningkatkan pemahaman antar kalimat. Dengan modifikasi ini, ALBERT secara signifikan mengurangi jumlah parameter dibandingkan dengan BERT, membuatnya lebih ringan dan lebih cepat untuk diimplementasikan.

- E_1, E_2, \dots, E_N : Merupakan representasi *embedding* dari input (kata atau token).
- Trm: Merupakan lapisan *transformer* yang bertanggung jawab untuk memproses *embedding* melalui *self-attention* dan *feed-forward network*.
- T_1, T_2, \dots, T_N : Merupakan *output* akhir yang digunakan untuk tugas downstream seperti klasifikasi, prediksi, atau lainnya.

| Model | | Parameters | Layers | Hidden | Embedding | Parameter-sharing |
|--------|---------|------------|--------|--------|-----------|-------------------|
| BERT | base | 108M | 12 | 768 | 768 | False |
| | large | 334M | 24 | 1024 | 1024 | False |
| | xlarge | 1270M | 24 | 2048 | 2048 | False |
| ALBERT | base | 12M | 12 | 768 | 128 | True |
| | large | 18M | 24 | 1024 | 128 | True |
| | xlarge | 60M | 24 | 2048 | 128 | True |
| | xxlarge | 235M | 12 | 4096 | 128 | True |

Gambar () membandingkan dua jenis model, yaitu *BERT* dan *ALBERT*, berdasarkan ukuran dan efisiensinya. *BERT* memiliki ukuran yang lebih besar, dengan jumlah parameter yang jauh lebih banyak, karena tidak menggunakan teknik *parameter sharing* (setiap layer menggunakan parameter yang berbeda). Di sisi lain, *ALBERT* dirancang untuk lebih efisien dengan jumlah parameter yang jauh lebih kecil karena menggunakan *parameter sharing* (parameter yang sama digunakan di beberapa layer).

| Model | | Parameters | SQuAD1.1 | SQuAD2.0 | MNLI | SST-2 | RACE | Avg | Speedup |
|--------|---------|------------|------------------|------------------|-------------|-------------|-------------|-------------|---------|
| BERT | base | 108M | 90.4/83.2 | 80.4/77.6 | 84.5 | 92.8 | 68.2 | 82.3 | 17.7x |
| | large | 334M | 92.2/85.5 | 85.0/82.2 | 86.6 | 93.0 | 73.9 | 85.2 | 3.8x |
| | xlarge | 1270M | 86.4/78.1 | 75.5/72.6 | 81.6 | 90.7 | 54.3 | 76.6 | 1.0 |
| ALBERT | base | 12M | 89.3/82.3 | 80.0/77.1 | 81.6 | 90.3 | 64.0 | 80.1 | 21.1x |
| | large | 18M | 90.6/83.9 | 82.3/79.4 | 83.5 | 91.7 | 68.5 | 82.4 | 6.5x |
| | xlarge | 60M | 92.5/86.1 | 86.1/83.1 | 86.4 | 92.4 | 74.8 | 85.5 | 2.4x |
| | xxlarge | 235M | 94.1/88.3 | 88.1/85.1 | 88.0 | 95.2 | 82.3 | 88.7 | 1.2x |

Pada gambar () merupakan bentuk dari tiap model dengan paramater yang berbeda – beda. Berikut penjelasan dari tiap model

1. Mode *A Lite BERT (ALBERT)*

- *Base* memiliki 12 juta parameter, jauh lebih kecil dari model *BERT base*, dan performa dari tipe ini menunjukkan efisiensi terhadap waktu yang dikeluarkan saat komputasi dan dengan waktu jauh lebih cepat 21.1 x
- *Large* memiliki 18 juta parameter, lebih kecil dari mode *BERT large*, dan performa dari tipe ini masih menunjukkan efisiensi dan hasil kinerja yang baik dan dengan waktu pemrosesan akan terus menurun sebesar 6.5 x. Hal ini karena jumlah parameter terus bertambah sesuai dengan ukuran yang digunakan.
- *Xlarge* memiliki 60 M menunjukkan bahwa performa pada tipe ini dengan kinerja yang baik dan dengan waktu pemrosesan akan melambah seiring bertambahnya parameter sebesar 2.4 x
- *Xxlarger* memiliki 235 M menunjukkan bahwa performa pada tipe ini dengan efisiensi yang lebih tinggi dalam pelatihan dengan waktu pemrosesan yang lebih lambat karena tipe ini memiliki parameter jauh lebih besar sehingga waktu pemrosesan sebesar 1.2 x

Parameter merujuk pada bobot (*weights*) dalam jaringan neural selama proses pelatihan model. Parameter merupakan model memproses data dan menghasilkan prediksi. Berikut parameter dalam model NLP *BERT* dan *ALBERT* mencakup

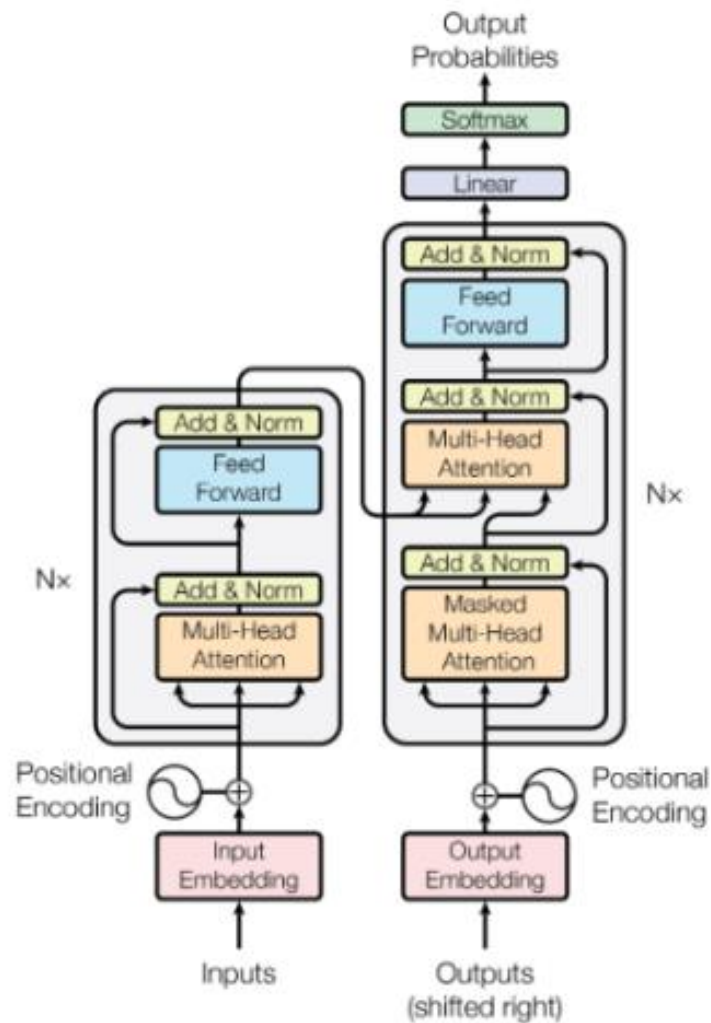
1. *Embedding layer* merupakan representasi vektor untuk setiap kata atau token dalam sebuah kosakata. Bobot tersebut akan memetakan kata menjadi representasi numerik dalam ruang dimensi tinggi.

2. *Layer transformer* merupakan paramter dalam mekanisme *self-attention*, *feed-forward networks* dan layer lainnya yang memproses representasi konteks antar kata
3. *Output layer* merupakan bobot yang digunakan untuk membuat prediksi seperti klasifikasi, analisis sentiemn, dan lainnya)

2.11 *Transormer*

Ttransformer merupakan sebuah arsitektur *deep learning* yang dirancang secara khusus untuk mendukung berbagai tugas dalam pemrosesan bahasa alami *Natural Language Processing* (NLP) di bidang kecerdasan buatan. Arsitektur ini menggunakan mekanisme perhatian (*attention mechanism*) yang memungkinkan pemrosesan teks secara paralel, sehingga mampu menghasilkan representasi kata yang mendalam dan sesuai konteks. *Transformer* dirancang untuk mempelajari hubungan antar kata atau entitas dalam suatu teks melalui dua komponen utama, yaitu *encoder* dan *decoder*. *Encoder* bertugas mengubah urutan simbol input menjadi representasi kontinu, sedangkan *decoder* menggunakan representasi tersebut untuk secara bertahap menghasilkan simbol *output*. Setiap lapisan pada *encoder* terdiri dari dua sub-lapisan penting, yakni mekanisme *self-attention multi-kepala* dan jaringan *feed-forward* yang sepenuhnya terhubung. Pada bagian *decoder*, struktur serupa diterapkan dengan tambahan satu lapisan yang menghubungkan hasil keluaran dari *encoder*. Untuk mendukung aliran informasi yang lebih baik, model ini memanfaatkan *residual connections* serta normalisasi lapisan. Berkat keunggulan arsitektur ini, *transformer* telah memberikan kontribusi

besar terhadap berbagai kemajuan di bidang NLP, seperti penerjemahan otomatis dan pemodelan teks (**Kutipan Transformer**). Berikut gambar arsitek *transformer*.



Gambar () menunjukkan alur dari arsitektur *transformer* mencakup dua bagian utama, yaitu *encoder* dan *decoder*, dengan langkah awal berupa *embedding* teks input. Kata-kata dalam teks diubah menjadi vektor *embedding*, lalu *Positional Encoding* ditambahkan untuk mempertimbangkan urutan kata karena *Transformer* tidak memiliki mekanisme bawaan untuk memahami posisi.

2.11.1 Encoder

Encoder merupakan beberapa lapisan yang masing-masing mencakup mekanisme perhatian multi-kepala untuk memahami hubungan antar kata, diikuti oleh jaringan *feed-forward* untuk memproses fitur lebih kompleks. *Residual connections* dan normalisasi diterapkan untuk menjaga stabilitas propagasi sinyal. Beberapa sub-komponen utama pada *encoder* diantaranya:

1. *Self-Attention Mechanism*

Merupakan proses dalam setiap lapisan yang dimana menghitung hubungan antar token dalam input. Setiap token memperhatikan semua token lainnya untuk memproses konteksnya. Formula *Self-Attention Mechanism* sebagai berikut

$$Attention(Q, K, V) = \text{Softmax} \left(\frac{Q \cdot K^T}{\sqrt{d_k}} \right) \cdot V$$

Penjelasan dari formula *Self-Attention Mechanism*:

- Query (Q) : Merupakan representasi token saat ini
- Key (K) : Merupakan token lainnya
- Value (V) : Merupakan informasi yang akan dibawa token
- $Q \cdot K^T$: Menghitung kesesuaian antar token
- $\sqrt{d_k}$: Merupakan normalisasi skala
- *Softmax* : Mengubah skor menjadi probabilitas

2. *Feed-Forward Neural Network*

Merupakan jaringan saraf dua lapisan yang bekerja secara terpisah pada setiap posisi token, dan bertujuan untuk meningkatkan representasi token.

3. *Layer Normalization*

Merupakan bagian yang bertujuan untuk menormalkan data di setiap lapisan, yang membantu kestabilan dan mempercepat konvergensi. Formula *Layer Normalization* sebagai berikut

$$\text{Norm}(x) = \frac{x - \mu}{\sigma} \cdot \gamma + \beta$$

Penjelasan dari formula *Layer Normalization*:

- x : Input vektor dari Dense Layer.
- μ : Rata-rata (mean) dari elemen-elemen vektor.
- σ : Standar deviasi dari elemen-elemen vektor.
- γ dan β : Parameter yang dipelajari, digunakan untuk mengembalikan skala output ke bentuk yang fleksibel.

4. *Residual Connection*

Merupakan inputan lapisan ditambahkan ke hasil *self-attention*, untuk memastikan aliran informasi tetap terjaga. Formula *residual connection* sebagai berikut

$$\text{Residual Output} = \text{Self-attention} + \text{Input to Layer}$$

2.11.2 *Decoder*

Decoder merupakan beberapa lapisan serupa tetapi dilengkapi dengan masked *multi-head attention* untuk memblokir informasi dari token masa depan, serta perhatian tambahan yang menyelaraskan output dengan hasil *encoder*. Hasil akhir dari *decoder* diproses melalui *softmax* untuk menghasilkan distribusi probabilitas kata, yang digunakan untuk membentuk output secara berurutan. Berikut Transfoermer

2.12 *Fine-Tuning*

Fine-tuning merupakan langkah untuk menyesuaikan model pembelajaran mesin yang sebelumnya telah melalui tahap pelatihan awal (*pre-trained model*), sehingga lebih relevan dengan tugas atau domain tertentu. Model ini umumnya telah mempelajari pola dan representasi data secara umum dari dataset skala besar, namun penyesuaian tambahan sering diperlukan agar model dapat menangkap karakteristik khusus dari data yang terkait dengan tugas baru. Dalam *fine-tuning* untuk analisis sentimen menggunakan fungsi *loss* dengan jenis *categorical cross-entropy loss*. *Categorical cross-entropy loss* merupakan bagian yang mengevaluasi kinerja model klasifikasi dalam analisis sentimen dengan membandingkan prediksi probabilitas model dengan label yang diinginkan. Fungsi ini menghitung perbedaan antara distribusi prediksi dan target, dengan memberikan penalti yang lebih besar untuk prediksi yang menyimpang jauh dari nilai sebenarnya. Rumusnya melibatkan perhitungan log probabilitas prediksi, di mana akurasi tinggi akan menghasilkan nilai *loss* yang lebih kecil (**Kutipan Fine-Tuning**). Formula *categorical cross-entropy loss* sebagai berikut;

$$Loss = \sum_{i=1}^N \sum_{j=1}^C y_{ij} \log(y_{ij})$$

Penjelasan dari formula *categorical cross-entropy loss*:

- y_{ij} adalah nilai one-hot encoded untuk label kelas yang benar.
- y^{ij} adalah probabilitas prediksi untuk kelas ke-j.
- C adalah jumlah kelas (dalam hal ini, 3 kelas: positif, negatif, netral).

2.13 *Support Vector Machine (SVM)*

Support Vector Machine (SVM) merupakan metode yang dirancang untuk melakukan klasifikasi dan regresi dengan memanfaatkan konsep hiperbidang untuk membagi data ke dalam dua kelompok. Prinsip utamanya adalah menentukan

hiperbidang yang ideal, yaitu yang memiliki margin terbesar untuk memisahkan kelas positif dan negatif. Untuk data non-linier, SVM menggunakan fungsi kernel yang memetakan data ke ruang berdimensi lebih tinggi agar menjadi linier dan lebih mudah dipisahkan (**Kutipan SVM**).

2.14 Evaluasi Model

Evaluasi model merupakan proses. Proses ini bertujuan untuk mengevaluasi dan menilai kinerja model yang telah dikembangkan guna mengidentifikasi kekuatan dan kelemahannya. Peneliti menggunakan metode Confusion Matrix sebagai alat untuk mengukur tingkat akurasi model dengan mengidentifikasi klasifikasi yang benar dan salah. Melalui metode ini, dapat dihitung metrik seperti accuracy, precision, recall, dan F1-score (**Kutipan Evaluasi Model**).

| | | True/Actual Class | |
|-----------------|-----------|---------------------|---------------------|
| | | Positive (P) | Negative (N) |
| Predicted Class | True (T) | True Positive (TP) | False Positive (FP) |
| | False (F) | False Negative (FN) | True Negative (TN) |
| | | P = TP + FN | N = FP + TN |

1. *Accuracy* merupakan metode evaluasi yang digunakan untuk menilai sejauh mana model mampu membuat prediksi yang benar dan salah dengan tingkat ketepatan tertentu. Berikut formula *accuracy*

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (2.1)$$

2. *Precision* merupakan metode evaluasi yang digunakan untuk mengukur sejauh mana ketepatan model dalam memprediksi kelas yang benar. Berikut formula *precision*

$$Precision = \frac{TP}{TP + FP} \quad (2.2)$$

3. *Recall* merupakan metode evaluasi yang digunakan untuk mengukur sejauh mana kemampuan model dalam mengidentifikasi nilai positif yang sesuai dengan data yang tersedia.

$$Recall = \frac{TP}{TP + FN} \quad (2.3)$$

4. *F1-score* merupakan metode evaluasi yang digunakan untuk menilai kinerja klasifikasi model dengan menggabungkan skor rata-rata antara *precision* dan *recall* yang telah diberi bobot.

$$F1-score = \frac{2 \times precision \times recall}{recall + precision} \quad (2.4)$$

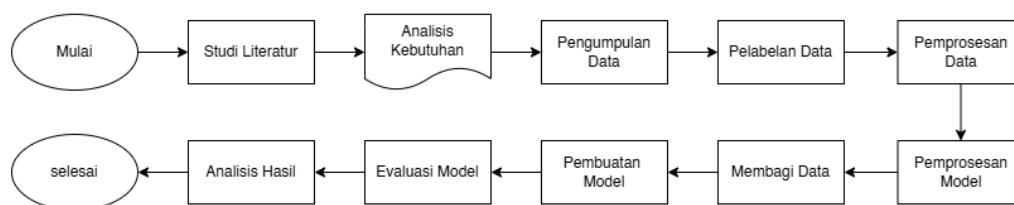
5. ROC-AUC merupakan batas klasifikasi yang menggambarkan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR). Berikut adalah rumus untuk TPR dan FPR.

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN} \quad (2.5)$$

III. METODE PELAKSANAAN

3.1 Tahapan Penelitian

Penelitian ini melibatkan alur diagram yang berperan penting sebagai panduan bagi penulis untuk menentukan alur dari penelitian agar berhasil dan terselesaikan. Penelitian ini melalui beberapa tahapan yang sistematis, seperti berikut pada gambar ().



3.2 Studi Literatur

Pada tahapan studi literatur, penulis mengumpulkan beberapa literatur sebagai pendukung dalam menyelesaikan penelitian ini meliputi literatur jurnal, artikel, dan referensi lainnya. Penulis memanfaatkan berbagai platform yang menyediakan berbagai jurnal yang relevansi terkait A Lite BERT sebagai metode ekstraksi fitur dan algoritma SVM sebagai penunjang dalam analisis sentimen serta platform media sosial X sebagai sumber data yang saya gunakan sebagai bahan sentimen.

3.3 Analisis Kebutuhan

Setelah melalui tahapan studi literatur, penulis melakukan ke tahapan selanjutnya yakni analisis kebutuhan terlebih dahulu untuk menunjang beberapa kebutuhan yang diperlukan dalam penelitian ini seperti perangkat keras dan perangkat lunak yang digunakan.

3.3.1 Spesifikasi Perangkat Keras

Pada penelitian ini penulis menggunakan perangkat keras berupa laptop pribadi dengan spesifikasi yang sesuai dengan kebutuhan pada penelitian ini. Laptop ini digunakan untuk mengimplementasikan dari semua tahapan penelitian mulai dari studi literatur sampai analisis hasil sebagai akhir penelitian. Berikut merupakan spesifikasi laptop yang saya gunakan pada tabel ().

| NO. | Perangkat Keras | Spesifikasi |
|-----|-----------------|--|
| 1 | Model Perangkat | Laptop Acer Aspire 5 A514 |
| 2 | Prosesor | Inter(R) Core(TM) i3-1005G1 CPU @ 1.20Hz |
| 3 | RAM | 8 GB |

3.3.2 Spesifikasi Perangkat Lunak

Pada penelitian ini penulis menggunakan beberapa kebutuhan perangkat lunak sesuai dengan kebutuhan pada penelitian ini. Perangkat lunak digunakan untuk mengimplementasikan dari semua alur tahapan pada penelitian mulai dari pemrosesan data sampai analisis hasil sebagai akhir penelitian. Berikut merupakan rincian perangkat lunak yang dibutuhkan pada tabel ().

| NO. | Perangkat Lunak | Spesifikasi |
|-----|--------------------|---|
| 1 | Tools | <ul style="list-style-type: none"> • Jupyter Notebook • Google Colabs |
| 2 | Bahasa Pemrograman | <ul style="list-style-type: none"> • Python |
| 3 | Library | <ul style="list-style-type: none"> • math • tweet harvest |

| | | |
|--|--|--|
| | | <ul style="list-style-type: none"> • transformers • torch • tenserflow • pandas • sklearn • numpy • scipy • matplotlib • scikit-learn |
|--|--|--|

3.4 Pengumpulan Data

Pada penelitian ini penulis menggunakan metode yakni crawling data untuk mendapatkan data primer berupa tweet dari pengguna media sosial X. Dataset ini merupakan hasil dari tweet dari pengguna yang merespon dari fenomena Kampanye sampai pelantikan presiden 2024. Dataset terdiri dari 10.000 data dan 16 atribut dalam format CSV. Rincian lengkap mengenai setiap atribut dapat dilihat pada tabel

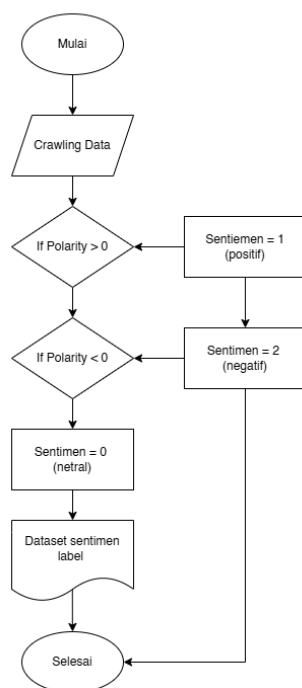
()

| No | Nama Atribut | Keteranagn |
|----|---------------------|---|
| 1 | Conversation_id_str | ID unik untuk sebuah percakapan |
| 2 | Created_at | Tanggal dan waktu tweet diposting |
| 3 | Favorite_at | Informasi waktu terakhir kali tweet di beri like oleh pengguna lain |
| No | Nama Atribut | Keterangan |

| | | |
|----|-------------------------|---|
| 4 | Full_text | Isi lengkap dari teks tweet, termasuk seluruh kata, link, hastag, dan mention |
| 5 | Id_str | ID unik untuk setiap tweet (bersifat individual) |
| 6 | Image_url | URL dari gambar yang disematkan dalam tweet |
| 7 | In_reply_to_screen_name | Nama pengguna (username) yang dituju jika tweet balasan (reply) tweet lain |
| 8 | Lang | Bahasa yang terdeteksi |
| 9 | Location | Lokasi geografis yang tertera pada profil |
| 10 | Quote_count | Jumlah tweet lain yang mengutip tweet ini |
| 11 | Reply_count | Jumlah balasan (reply) yang diberikan ke tweet ini |
| 12 | Retweet_count | Jumlah retweet yang diterima oleh tweet ini |
| 13 | Tweet_url | URL langsung ke halaman tweet tersebut di tweeter |
| 14 | User_id_str | ID unik pengguna yang memposting tweet tersebut |
| 15 | Username | Nama pengguna (username) dari akun yang memposting tweet |

3.5 Pelabelan Data

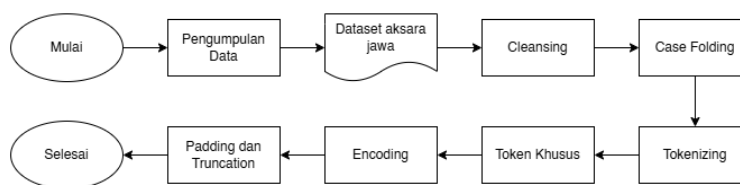
Data yang didapatkan dari *crawling data* akan dilanjutkan ke tahapan pelabelan data. Pada penelitian ini data diberi label atau kelas dengan 3 label (positif, negatif, netral). Positif menandakan bahwa opini tersebut mendukung kandidat, sedangkan negatif pertanda bahwa opini berbentuk seperti kritik atau isu – isu yang mengacu pada polarisasi, untuk netral tidak menandakan berupa opini negatif maupun positif. Dalam pelabelan data secara *semi-automatic labelling* dengan *lexicon based* untuk analisis sentimen



Pada gambar () menjelaskan dataset yang dihasilkan *crawling data* akan diberi label berdasarkan jumlah polaritas dari setiap kalimat yang ada pada dataset. *Lexicon based* memberikan sebuah nilai pada kata yang mengandung polaritas dengan mengandalkan kamus kata (*lexicon*) yang berisikan daftar kata atau frasa yang sudah diberi nilai polaritas. Sehingga dapat menentukan kalimat tersebut label sentimen dari kalimat tersebut.

3.6 Pemrosesan Data

Data yang didapatkan dari tahapan sebelumnya yang sudah diberi label akan dilanjutkan ke tahapan pemrosesan data. Pada penelitian ini data yang didapatkan masih bersifat mentah belum siap untuk di analisis. Oleh karena itu, sebelum lanjut ke tahapan selanjutnya data ini akan dilakukan pemrosesan data terlebih dahulu. Tujuan dari tahapan ini adalah untuk mendapatkan data yang siap dibutuhkan dalam analisis lebih lanjut dan mendapatkan analisis sesuai dengan penelitian. Alur tahapan pemrosesan data dapat dilihat pada gambar ()



Berikut penjelasan dari setiap alur tahapan pemrosesan data yang terdapat pada gambar ():

1. *Cleansing* dilakukan untuk menghilangkan beberapa atribut yang tidak dibutuhkan serta menghilangkan baik tanda baca, *emoticon*, url dan lain - lain. Memastikan tidak ada duplikasi data, dan *missing value*. Dalam tahapan ini hanya menggunakan atribut yang dibutuhkan seperti `full_text`, `username`, `created_at` dan dari atribut yang sudah dipilih akan dilanjutkan perubahan nama atribut menyesuaikan dengan penelitian ini terkait analisis sentimen dari data teks berupa opini masyarakat. Tujuan dari tahapan ini agar mendapatkan data yang bersih dan siap digunakan untuk mempermudah dalam analisis lebih lanjut.
2. *Case folding* dilakukan untuk mengonversi semua huruf dalam teks menjadi huruf kecil (*lowercase*). Tujuan dalam tahapan ini memastikan bahwa teks

dalam berbagai format penulisan memiliki format yang seragam dan konsisten, menghilangkan perbedaan yang disebabkan oleh huruf kapital, mengurangi dimensi dalam representasi data teks. Tahapan ini mempermudah penulis untuk meningkatkan efisiensi dan mempermudah dalam analisis sentimen dari data teks opini masyarakat. Berikut contoh tahapan *case folding* “Saya sukA BelAjAr Membaca” → “saya suka belajar membaca”

3. *Tokenizing* dilakukan untuk merujuk pada proses pemecahan teks menjadi bagian – bagian lebih kecil, seperti kata, sub – kata. Dalam tahapan ini menggunakan alat dari metode *A Lite BERT* yaitu *tokenizer* dari *WordPiece* yang bertujuan untuk merepresentasikan subword untuk menangani di luar kosatakata (*out of vocabulary*). Berikut contoh *tokenizing* “saya suka belajar membaca” → [“saya”, “suka”, “”, “bel”, “” ##ajar” “mem”, “##baca”]

- Pada tanda “##” menunjukkan bahwa token tersebut adalah bagian dari kata sebelumnya. ##ajar adalah kelanjutan dari bel untuk membentuk kata "belajar". Dan ##baca adalah kelanjutan dari mem untuk membentuk kata "membaca".

4. Token khusus dilakukan untuk merancang dalam membantu memahami struktur teks, konteks, dan relasi antar bagian teks. Token khusus berperan penting dalam klasifikasi dan analisis sentimen. Token khusus yaitu berupa sebagai berikut:

- *Classification token* [CSL] tujuan pemberian token ini untuk menunjukkan awal dari sebuah input atau kalimat dalam tugas klasifikasi teks maupun analisis sentimen. Penggunaan token ini

sebagai ringkasan untuk memungkinkan model untuk mengerti konteks keseluruhan kalimat.

- *Seperation token* [SEP] tujuan pemberian token ini untuk memisahkan bagian – bagian dalam input atau kalimat. Biasanya untuk memisahkan dua kalimat dalam satu input. Penggunaan token ini untuk memisahkan dua kalimat agar memastikan model tahu satu bagian berakhir dan bagian dimulai

Berikut contoh dari tahapan token khusus [“saya”, “suka”, “, “bel”, ” ##ajar” “mem”, “##baca”] → “[CLS]”, “saya”, “suka”, “bel”, “##ajar”, “mem”, “##baca”, “[SEP]”]

5. *Encoding* dilakukan untuk mengubah dari data hasil tokenezing yang sudah diberi inputan token khusus menjadi representasi numerik berupa token ID berupa bilangan bulat (*integer*). Proses ini dilakukan dengan mencocokkan setiap token dengan indeksnya di kosatkata (*vocabulary*) model.

- Kosakata (*vocabulary*) terdiri dari tabel yang berisi daftar semua token yang dikenali oleh model beserta ID-nya. Setiap kosakata diberikan urutan ID berdasarkan posisinya dalam kosakata. Urutan berdsakaan struktur kosakata, untuk urutan token khusus terlebih dulu [PAD], [CLS], [SEP], [MASK]. Sebagai contoh pada tahapan *encoding*

[PAD] → 0, [CLS] → 1, [SEP] → 2, [MASK] → 3, [CLS] → 1, “saya” → 4, “suka” → 5, “bel” → 6, “##ajar” → 7, “mem” → 8, “##baca” → 9, [SEP] → 2

Gabungan ID token menjadi daftar: [1,4, 5, 6, 7, 8, 9,2]

Penentuan ID token seperti [1,4, 5, 6, 7, 8, 9,2] tersebut karena dalam algoritma tokenisasi *wordpiece* tidak acak, tetapi berdasarkan struktur kosakata yang dibuat. ID hanya satu angka atau bertambah secara berurutan karena itu digunakan oleh model.

6. *Padding* dan *truncation* dilakukan untuk menambahkan token tertentu [PAD] pada akhir atau awal kalimat untuk membuat panjang input konsisten. Tujuan *padding* dilakukan untuk memastikan bahwa semua input memiliki panjang yang konsisten, terutama batch input diberikan ke model. *Padding* umumnya digunakan pada *A Lite BERT* atau *BERT* 512 token. Akan tetapi penentuan panjang maksimum atau batasan sebuah token menyesuaikan dengan kebutuhan token bisa berupa 64, 128, 256, 512 itu semua tergantung kebutuhan sebagai implementasi *padding* [PAD]:

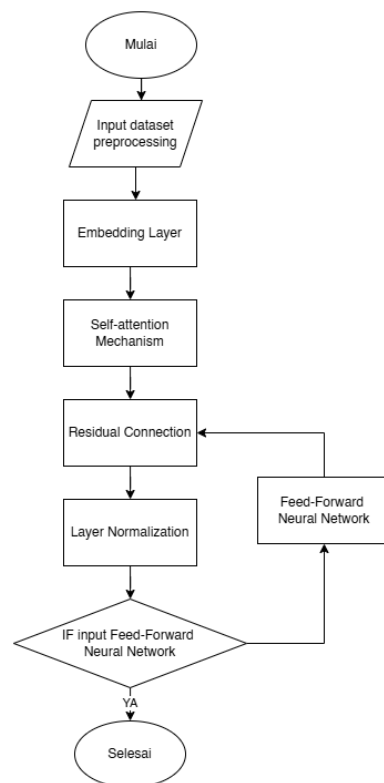
- Kalimat: ["[CLS]", "saya", "suka", "bel", "##ajar", "mem", "##baca", "[SEP]"] dengan encoding [1,4, 5, 6, 7, 8, 9,2]
- Hasil *padding* [PAD] : [1, 4, 5, 6, 7, 2, 0, 0, ..., 0] dimana pemberian token ID [PAD] → 0 sampai mencapai total 512 token

Untuk *truncation* dilakukan dalam memotong teks (*sequence* token) yang panjangnya melebihi panjang maksimum yang ditentukan misal batas token *padding* 512 token. Jika melebihi dari batas yang sudah ditentukan maka token selanjutnya akan dipotong. Sebagai contoh untuk batas token *padding* 6 toke

- ["[CLS]", "saya", "suka", "bel", "##ajar", "mem", "##baca", "[SEP]"]
→ [1,4, 5, 6, 7, 8, 9,2] = 8
- ["[CLS]", "saya", "suka", "bel", "##ajar", "[SEP]"] → [1, 4, 5, 6, 7, 2] Menghilangkan kata "mem", "##baca"

3.7 Transformer Model A Lite BERT

Tahapan ini melibatkan dataset yang didapatkan dari hasil tahapan pemrosesan data yakni berbentuk ID token, dilanjutkan ke tahapan ini dengan tujuan untuk mengubah ID token menjadi representasi numerik dengan model transformer A Lite BERT, representasi dilakukan melalui *embedding layer*, *self-attention*, dan *feed-forward neural network*. Berikut alur tahapan pemrosesan model dapat dilihat pada gambar ().

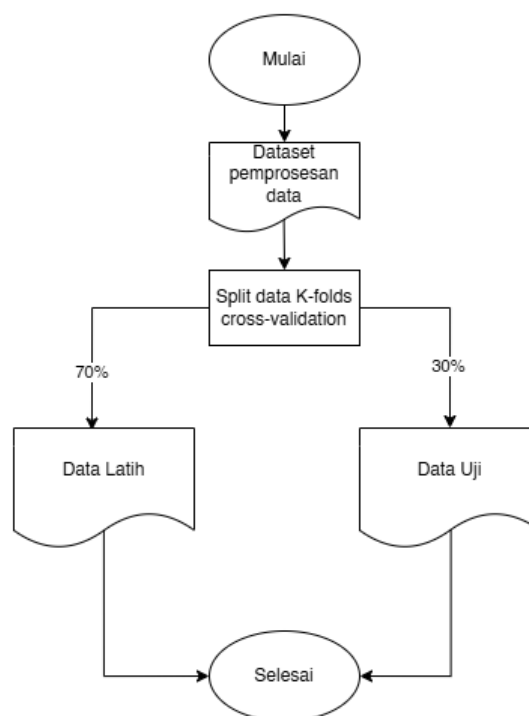


Berdasarkan gambar (), tahapan ini meliputi *embedding layer* yang mengubah token ID menjadi vektor *embedding* dengan mencakup informasi dari tipe token *type embedding* dan representasi dasar kata (*word embedding*). Hasil dari *embedding layer* lanjut ke tahapan *self-attention* untuk memahami hubungan antar token dalam kalimat. *Self-attention* menangkap konteks global dengan menghitung

attention antar token. Hasil dari *self-attention* lanjut ke tahapan *feed-forward neural network* untuk menangkap fitur yang lebih kompleks dengan dimensi baru menggunakan transformasi linier, memungkinkan model menangkap pola lebih kompleks dari data yang digunakan.

3.8 Membagi Data (*Split Data*)

Membagi Data (*split data*) tahapan ini penting pada penelitian penulis karena bertujuan untuk memisahkan dataset menjadi beberapa bagian terpisah dengan tujuan yang berbeda. Berikut alur *split data* pada gambar ()

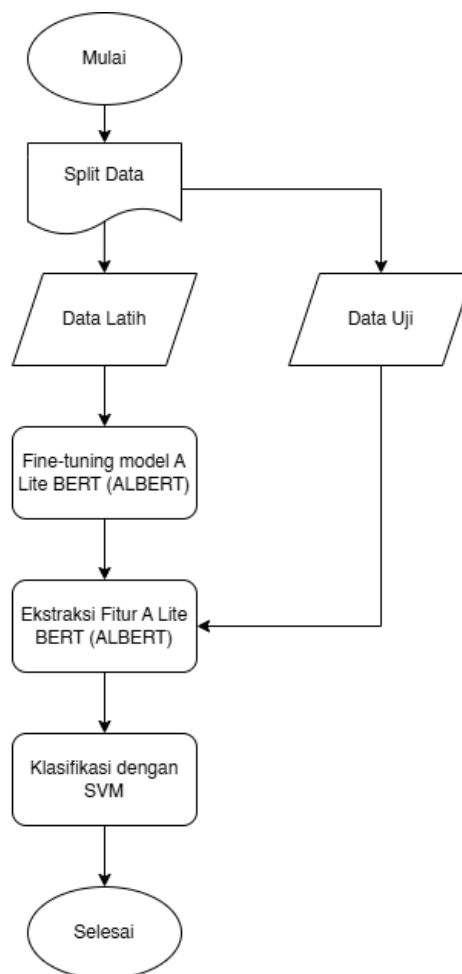


Data yang sudah melewati pemrosesan model akan melalui tahapan pembagian data dengan menggunakan metode *K-fold cross – validation* yaitu metode yang digunakan untuk membagi data menjadi beberapa bagian (*folds*) dan secara bergantian menggunakan setiap bagian sebagai data latih, dan data uji. Pada penelitian ini penulis membagi dataset 10.000 menjadi tiga bagian dengan porsi

yang sudah ditentukan sebesar 70% data latih yang berguna untuk melatih model, dan data uji 30% digunakan untuk mengevaluasi hasil dari model yang sudah dilatih. Metode ini penting untuk memastikan bahwa model tidak mengalami masalah *overfitting* atau *underfitting*, sehingga mampu menghasilkan generalisasi yang baik saat digunakan pada data nyata.

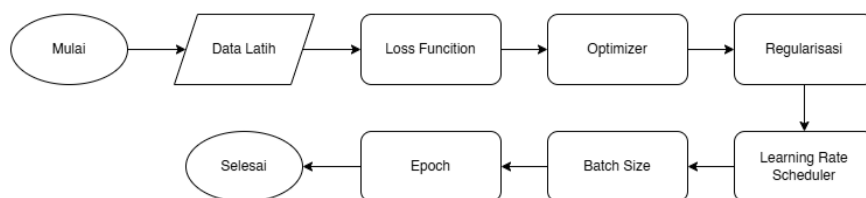
3.9 Pembuatan Model

Data yang sudah dibagi menjadi tiga jenis di lanjutkan ke tahapan pembuatan model. Tujuan tahapan ini adalah mempersiapkan model bekerja pada data yang sudah melewati beberapa tahapan untuk di analisis terkait sentimen dari data yang sudah di labeli. Proses pembuatan model dapat dilihat pada gambar ()



3.9.1 *Fine-Tuning A Lite BERT*

Fine-tuning dalam tahapan ini sebagai melatih kembali model yang sudah di latih sebelumnya untuk memanfaatkan pengetahuan yang sudah diperoleh model dari tahapan pre-training untuk mempercepat kinerja pada data, dapat dilihat pada gambar ().



Berdasarkan gambar (), dengan memasukkan data latih nantinya akan dilakukan pengukuran kesalahan antar prediksi model dan label sebenarnya melalui tahapan *loss function* yang hasilnya nanti akan dilanjutkan untuk memperbarui bobot model berdasarkan gradien yang dihitung melalui tahapan *optimizer*. Dari hasil tersebut dilanjutkan untuk membatasi kompleksitas model untuk mencegah overfitting yang di tahapan regularisasi yang nanti hasilnya akan dilanjutkan untuk mengatur *learning rate* agar proses pembelajaran lebih stabil yang dimulai kecil dan perlahan menurun mendekati nol. Setelah proses tahapan ini dilanjutkan dengan menentukan jumlah data yang diproses dalam satu iterasi, untuk setiap batch menghasilkan satu langkah pembaruan bobot berdasarkan *loss* dan gradien, setelah melalui pembobotan dari setiap batch dilanjutkan dengan melakukan iterasi penuh melalui seluruh dataset, dimana setelah satu epoch

selesai, *validation loss* di ukur untuk memantau performa model, model akan kembali ke langkah 1 untuk memulai iterasi berikutnya (epoch berikutnya) menggunakan data yang untuk mencapai model yang memadai.

3.9.2 Ekstraksi Fitur A Lite BERT

Ekstraksi Fitur menggunakan model A Lite BERT dalam tahapan ini sebagai mengubah representasi numerik yang lebih mudah di pahami oleh model klasifikasi dengan menggunakan *hidden state* Dapat dilihat pada gambar ()



Gambar

Berdasarkan gambar (), dengan data dari hasil *fine-tuning* dilakukan proses tahapan *hidden states* dari token [CSL] untuk mengandung informasi keseluruhan tentang teks, yang mencakup makna sematik dan konteks dari kalimat yang dianalisis dengan begitu representasi numerik ini menghasilkan vektor fitur yang representatif dari teks yang akan digunakan untuk proses klasifikasi dan analisis sentimen.

3.9.3 Klasifikasi dengan SVM

Tahapan ini menggunakan data hasil dari ekstraksi fitur yang sudah memiliki representasi numerik yang menghasilkan berupa vektor fitur yang representatif dari teks seta sematik dan memahami konteks kalimat. Untuk data latih yaitu berguna untuk melatih model SVM untuk data pelatihan yang sudah memiliki label positif, negatif atau netral. Dan untuk data uji digunakan untuk

menguji model yang sudah di latih menggunakan data latih, dan SVM mengklasifikasikan data uji berdasarkan posisis relatif terhadap hyperplane.

3.10 Evaluasi Model

Setelah melewati tahapan klasifikasi menggunakan SVM, evaluasi model perlu dilakukan agar memastikan hasil dari tahapan sebelumnya menghasilkan berupa akurasi, presisi, *recall*, *f1-score*, dengan presentasi atau kinerja yang baik dengan menggunakan *confusion matrix*. Dari tahapan ini mengukur presentasi prediksi dari data uji seberapa banyak prediksi yang benar sesuai labe tertentu dibandingkan semua prediksi pada label tersebut. *Recall* menilai sejauh mana model mampu mendeteksi semua sampel dari kelas tertentu, dan F1-score memberikan keseimbangan antara presisi dan recall, yang sangat berguna ketika data tidak seimbang. Confusion matrix menyajikan informasi tentang distribusi prediksi benar dan salah untuk setiap label.

3.11 Analisis Hasil

Setelah model di evaluasi, langkah selanjutnya pada tahapan ini yaitu melakukan analisis hasil evaluasi model SVM menggunakan metode ekstraksi fitur A Lite BERT untuk memahami kinerja model dan mengidentifikasi kekuatan serta kelemahannya. Dalam analisis ini, penulis memeriksa metrik evaluasi seperti akurasi, presisi, recall, F1-score, dan confusion matrix untuk setiap label (positif, negatif, netral). nalisis ini tidak hanya memberikan gambaran performa model secara keseluruhan tetapi juga membantu menemukan aspek spesifik yang memerlukan perbaikan.

DAFTAR PUSTAKA