ANALISIS SENTIMEN MENGENAI BERITA PEMBERHENTIAN GENOSIDA MENGGUNAKAN DATA X DENGAN ALGORITMA BERT DAN *NAÏVE BAYES CLASSIFIER*

(Skripsi)

Oleh

TASYA CYNTHIA MONICA LOVELINDRA
1955061004

FAKULTAS TEKNIK
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

ANALISIS SENTIMEN MENGENAI BERITA PEMBERHENTIAN GENOSIDA MENGGUNAKAN DATA X DENGAN ALGORITMA BERT DAN *NAÏVE BAYES CLASSIFIER*

Oleh

TASYA CYNTHIA MONICA LOVELINDRA
1955061004

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA TEKNIK

Pada

Program Studi Teknik Informatika
Jurusan Teknik Elektro
Fakultas Teknik Universitas Lampung

FAKULTAS TEKNIK
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

ABSTRAK

ANALISIS SENTIMEN MENGENAI BERITA PEMBERHENTIAN GENOSIDA MENGGUNAKAN DATA X DENGAN ALGORITMA BERT DAN *NAÏVE BAYES CLASSIFIER*

Oleh

TASYA CYNTHIA MONICA LOVELINDRA

Genosida merupakan kejahatan luar biasa yang bertujuan untuk memusnahkan kelompok bangsa, ras, etnis, atau agama tertentu. Isu ini sering menjadi sorotan masyarakat dunia, termasuk di media sosial X, tempat banyak pengguna menyampaikan opini terkait peristiwa genosida seperti genosida Israel terhadap Palestina. Namun, opini tersebut belum dapat diketahui kecenderungan sentimennya, apakah positif, negatif, atau netral. Oleh karena itu, penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap isu genosida dengan memanfaatkan algoritma *Naïve Bayes Classifier dan Bidirectional Encoder Representations from Transformers* (BERT). Penelitian ini mengacu pada metode CRISP-DM yang meliputi tahapan *business understanding, data understanding,* data *preparation, modelling, dan evaluation*. Data dikumpulkan dari media sosial X melalui API, kemudian diproses melalui tahap *preprocessing* dan ekstraksi fitur menggunakan TF-IDF untuk *Naïve Bayes* serta tokenisasi untuk BERT. Hasil penelitian menunjukkan bahwa *Naïve Bayes Classifier* memperoleh akurasi sebesar 81%, sedangkan BERT memperoleh akurasi sebesar 73%. Walaupun *Naïve Bayes* memiliki akurasi yang lebih tinggi, BERT mampu menangkap konteks kalimat yang lebih kompleks dan menunjukkan performa yang lebih konsisten pada setiap kelas sentimen, sehingga tetap relevan diterapkan untuk analisis sentimen yang membutuhkan pemahaman konteks mendalam. Dengan demikian, kedua algoritma memiliki keunggulannya masing-masing, dan pemilihannya dapat disesuaikan dengan kebutuhan analisis.

Kata kunci: Genosida, Analisis Sentimen, Media Sosial X, *Naïve Bayes*, BERT

ABSTRACT

Sentiment Analysis on News About the Termination of Genocide Using X Data with BERT and Naïve Bayes Classifier Algorithms

By

TASYA CYNTHIA MONICA LOVELINDRA

Genocide is an extraordinary crime aimed at eliminating specific national, racial, ethnic, or religious groups. This issue often becomes a global concern, including on the social media platform X, where many users express their opinions about genocide related events, such as the genocide committed by Israel against Palestine. However, the sentiment tendency of these opinions whether positive, negative, or neutral cannot be identified directly. Therefore, this study aims to analyze public sentiment toward genocide issues by employing the Naïve Bayes Classifier algorithm and the Bidirectional Encoder Representations from Transformers (BERT). The research follows the CRISP-DM methodology, which includes the stages of business understanding, data understanding, data preparation, modelling, and evaluation. Data were collected from the X platform through an API, then preprocessed and transformed using TF-IDF for Naïve Bayes and tokenization for BERT. The results show that the Naïve Bayes Classifier achieved an accuracy of 81%, while BERT obtained an accuracy of 73%. Although Naïve Bayes produced a higher accuracy, BERT demonstrated a stronger ability to capture complex contextual information and exhibited more consistent performance across sentiment classes, making it highly relevant for sentiment analysis tasks that require deep contextual understanding. Thus, both algorithms possess their respective strengths, and their selection can be adjusted based on analytical needs.

Keywords: Genocide, Sentiment Analysis, Social Media X, Naïve Bayes, BERT

Judul Skripsi	:	**ANALISIS SENTIMEN MENGENAI BERITA PEMBERHENTIAN GENOSIDA MENGGUNAKAN DATA X DENGAN ALGORITMA BERT DAN *NAÏVE BAYES CLASSIFIER.***
Nama Mahasiswa	:	Tasya Cynthia Monica Lovelindra
Nomor Pokok Mahasiswa	:	1955061004
Program Studi	:	Teknik Informatika
Jurusan	:	Teknik Elektro
Fakultas	:	Teknik

MENYETUJUI

1. Komisi Pembimbing

Pembimbing Utama	Pembimbing Pendamping
Dr. Ir. M. Komarudin, S.T.,M.T	**Ir. Titin Yulianti, S.T.,M.Eng**
NIP. 196812071997031006	NIP.198807092019032015

2. Mengetahui

Ketua Jurusan	Ketua Program Studi
Teknik Elektro	Teknik Informatika
Herlinawati, S.T.,M.T.	**Yessi Mulyani, S.T.,M.T.**
NIP. 197103141999032001	NIP. 197312262000122001

MENGESAHKAN

1. Tim Penguji

Ketua	:**Dr. Ir. M. Komarudin, S.T.,M.T**	.....
Sekretaris	:**Ir. Titin Yulianti, S.T ., M.Eng**	.....
Penguji	: **Wahyu Eko S, S.T ., M.Sc**	.....

2. Dekan Fakultas Teknik

Dr. Eng. Ir. Helmy Fitriawan, S.T., M.Sc.
NIP. 19750928 200112 1 002

Tanggal Lulus Ujian Skripsi : 04 Desember 2025

SURAT PERNYATAAN

Saya yang bertanda tangan dibawah ini, menyatakan bahwa skripsi saya dengan judul "Analisis Sentimen Mengenai Berita Pemberhentian Genosida Menggunakan Data X Dengan Algoritma BERT dan *Naïve Bayes Classifier* " dibuat oleh saya sendiri. Semua hasil yang tertuang dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung.

Apabila di kemudian hari terbukti bahwa skripsi ini merupakan salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan hukum atau akademik yang berlaku.

Bandar Lampung, 04 Desember 2025
Pembuat Pernyataan,

Tasya Cynthia Monica Lovelindra
NPM. 1955061004

RIWAYAT HIDUP

Penulis bernama Tasya Cynthia Monica Lovelindra lahir pada tanggal 17 November 2000. Penulis merupakan anak sulung dari pasangan Bapak Indra Gunawan, S.E dan Ibu Lismayani Falsa, S.E. Penulis mengawali pendidikan di TK Pratama Bandar Lampung. Selanjutnya pada tahun 2006 sampai 2012, penulis melanjutkan pendidikan di SDN 2 Rawa Laut Bandar Lampung. Kemudian pada tahun 2012 sampai 2015 melanjutkan pendidikan di SMP Kartika II-2 (Persit) Bandar Lampung, lalu melanjutkan ke SMA Negeri 2 Bandar Lampung dan tamat pada tahun 2018. Kemudian pada tahun 2019 penulis diterima sebagai mahasiswa baru pada Program Studi Teknik Informatika, Jurusan Teknik Elektro, Fakultas Teknik, Universitas Lampung melalui jalur SMMPTN.

Selama menjadi mahasiswa, penulis aktif dalam beberapa kegiatan, antara lain Menjadi anggota biasa Himpunan Mahasiswa Teknik Elektro Universitas Lampung, Departemen Sosial dan Kewirausahaan, Divisi Sosial pada periode 2019/2020. Kemudian pada periode 2020/2021 menjadi Bendahara Umum Himpunan Mahasiswa Teknik Elektro Universitas Lampung. Pada tahun 2021 Mengikuti program Studi Independen Bersertifikat Kampus Merdeka dari PT. Microsoft Indonesia dengan program Data dan *Artificial Intelligence*. Kemudian pada tahun 2022 mengikuti program Studi Independen Bersertifikat Kampus Merdeka dari Orbit Future Academy dengan program *Mastery* AI. Penulis Melaksanakan Kuliah Kerja Nyata (KKN) pada bulan Januari sampai dengan Febuari 2022 di Kelurahan Bumi Kedamaian, Kecamatan Kedamaian, Kota Bandar Lampung. Kemudian pada bulan Juli 2022 penulis melakukan Praktik Kerja Lapangan di PT Kazee Digital Indonesia pada divisi Data *Science*.

MOTTO

"Sesungguhnya bersama kesulitan ada kemudahan. Maka apabila kamu telah selesai (dari sesuatu urusan), tetaplah bekerja keras (untuk urusan yang lain). Dan hanya kepada Tuhanmulah kamu berharap"

(Q.S Al-Insyirah: 6-8)

"Allah tidak membebani seseorang melainkan sesuai dengan kesanggupannya"

(Al Baqarah 286)

"Semua sudah tertulis, menawarlah dengan doa"

(Hearty Servi)

"Selesaikan apa yang sudah kamu mulai. Tuntaskan apa yang sudah kamu ambil. Mari nikmati proses ini. Mari berjalan meski pelan."

(Alkana)

"Tidak ada mimpi yang terlalu tinggi dan tidak ada mimpi yang patut diremehkan. Lambungkan setinggi yang kau inginkan dan gapailah dengan selayaknya yang kau harapkan".

(Maudy Ayunda)

"*Life can be heavy, especially if you try to carry it all at once. Part of growing up and moving into new chapters of your life is about catch and release"*

(Taylor Swift)

"it will pass"

PERSEMBAHAN

Sujud syukur kupersembahkan kepada Allah SWT, Tuhan Yang Maha Agung danMaha Tinggi. Atas takdirmu saya bisa menjadi pribadi yang berpikir, berilmu, beriman dan bersabar. Semoga keberhasilan ini menjadi satu langkah awal untukmasa depanku, dalam meraih cita-cita.

Kupersembahkan karya ilmiah ini teruntuk:

"Untuk Papaku sayang yang berada di surganya Allah.
Aku tumbuh dari doa-doa papa, dan aku berdiri sampai hari ini karena cinta dan pengorbanan papa yang tak ternilai. Walau tidak bisa lagi merayakan pencapaian ini bersama, aku selalu percaya papa melihatku dari jauh dan tetap menjadi alasanku bertahan. Pa, maaf apabila menunggu begitu lama aku sudah berada di titik akhir dari apa yang kamu harapkan. Gelar ini yang kamu nantikan aku persembahkan untuk papa dan aku sudah menepati janjiku. Lihatlah aku dari sana dan berbahagialah disana atas semuanya."

"Untuk Mamaku tercinta, yang selalu berjuang tanpa henti demi memberikan yang terbaik untukku. Terima kasih atas kasih sayang, doa, dan dorongan yang tak pernah putus di setiap langkah hidupku. Semua pengorbanan Mama tak akan pernah mampu terbalas hanya dengan beberapa kalimat di halaman persembahan ini. Aku sayang Mama, tetaplah hidup lebih lama mendampingi aku dan adik-adik."

"untuk dedek Digo dan Adek Farel tersayang, yang selalu memberikan dukungan dan doa yang terbaik untukku. Serta juga terima kasih kepada diriku yang telah berjuang sampai akhir, walau penuh dengan halangan dan rintangan. Semoga segala harapan dan impianmu dapat segera terwujud.

SANWACANA

Segala puji dan syukur penulis panjatkan ke hadirat Allah SWT, Tuhan semesta alam yang Maha Pengasih dan Maha Penyayang. Berkat rahmat, taufik, dan hidayah-Nya, penulis akhirnya dapat menyelesaikan skripsi berjudul **"Analisis Sentimen Mengenai Berita Pemberhentian Genosida Menggunakan Data X Dengan Algoritma BERT dan *Naïve Bayes Classifier*".** Penyusunan skripsi ini merupakan salah satu syarat untuk memenuhi kelulusan dan memperoleh gelar Sarjana Teknik pada Program Studi Teknik Informatika, Jurusan Teknik Elektro, Fakultas Teknik, Universitas Lampung. Dalam proses penelitian hingga penyusunan laporan ini, penulis menerima begitu banyak bantuan berupa bimbingan, ide, fasilitas, serta motivasi dari berbagai pihak. Untuk itu, dengan penuh rasa hormat penulis mengucapkan terima kasih kepada:

1. Bapak Dr. Eng. Ir. Helmy Fitriawan, S.T., M.Sc. selaku Dekan Fakultas Teknik Universitas Lampung.
2. Ibu Herlinawati, S.T., M.T. selaku Ketua Jurusan Teknik Elektro Universitas Lampung.
3. Ibu Yessi Mulyani, S.T., M.T. selaku Ketua Program Studi Teknik Informatika Universitas Lampung yang telah memberikan nasihat, arahan, saran, dan motivasi kepada penulis.
4. Bapak Dr. Ir. M. Komarudin, S.T., M.T, selaku Pembimbing Utama yang telah bersedia meluangkan waktu untuk memberikan pengarahan dan bimbingan dalam pembuatan skripsi ini hingga selesai.
5. Ibu Ir. Titin Yulianti, S.T., M.Eng, selaku Pembimbing Pendamping yang telah bersedia meluangkan waktu untuk memberikan pengarahan dan bimbingan dalam pembuatan skripsi ini hingga selesai.

6. Bapak Wahyu Eko S, S.T., M.Sc, selaku Penguji yang telah memberikan nasihat, arahan, saran, dan motivasi kepada penulis.
7. Seluruh Dosen Program Studi Teknik Informatika yang telah membagikan ilmunya kepada penulis.
8. Seluruh jajaran staf administrasi atas bantuannya dalam menyelesaikan urusan administrasi di Jurusan Teknik Elektro Universitas Lampung.
9. Sahabat penulis Citra Mutiara Putri, Amirah Ghina Salsabila, Dhira Atika, Olivia Amarezha, dan Nediyan Fitri Anissa. Terima kasih atas segala dukungan dan semangat yang diberikan ke penulis.
10. Sahabat-sahabat penulis Atiqah Hanifah Shalihah, Alya Nurul Fakhira, Fiona Yovita Syafri, Aurellia Salma Fertiyan, Azzahra Agitha dan seluruh teman-teman Teknik Informatika dan Teknik Elektro Angkatan 2019 atas dukungan yang telah diberikan selama menempuh studi di Program Studi Teknik Informatika Universitas Lampung.
11. Teruntuk Yesaya Abraham Sitanggang terima kasih telah hadir diwaktu yang tepat, terima kasih telah memerankan karakter Mas Trian, kemudian kepada Zara Adisty terima kasih hadir sebagai adila yang sangat memotivasi agar selalu sabar serta Beri Cinta Waktu yang lainnya terimakasih. Kemudian Actor Leeminho yang sedari kecil selalu buat penulis semangat menjalankan hidup.

Semoga Allah SWT membalas segala kebaikan dan bantuan yang telah diberikan kepada penulis. Penulis berharap skripsi ini dapat membawa manfaat bagi para pembaca serta menjadi bekal penulis dalam mengembangkan dan menerapkan ilmu yang telah dipelajari.

Bandar Lampung, 04 Desember 2025
Penulis,

Tasya Cynthia Monica.L
NPM. 1955061004

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

I. PENDAHULUAN

1.1 Latar Belakang

Genosida adalah wujud dari kejahatan yang melibatkan upaya pemusnahan suatu etnis, budaya, atau kelompok tertentu, termasuk kelompok politik yang sulit diidentifikasi sehingga dapat menimbulkan persoalan di tingkat internasional. Konvensi Genosida 1948 menjelaskan bahwa genosida adalah perbuatan yang dijalankan dengan tujuan memusnahkan seluruh atau sebagian kelompok suatu bangsa, ras, etnis, atau agama. Rumusan ini lalu diadopsi berdasarkan Statuta *International Criminal Court* (ICC) serta Undang-Undang Nomor 26 Tahun 2000 tentang Pengadilan HAM, yang menegaskan bahwa kelompok bangsa, ras, maupun etnis memiliki identitas, ciri, dan tradisi turun-temurun yang membedakan mereka dari kelompok lain. Dalam hukum pidana internasional, genosida dikategorikan sebagai kejahatan kelas berat dan termasuk perbuatan yang tidak diizinkan, sebagaimana tercantum dalam Konvensi Genosida 1948, Statuta *International Criminal Tribunals for the Former Yugoslavia* (ICTY), Statuta *International Criminal Tribunals for Rwanda* (ICTR), serta Statuta Roma 1998 yang menegaskan genosida sebagai kejahatan paling serius yang menjadi perhatian seluruh komunitas internasional. Di Indonesia, Pasal 7 Undang-Undang Pengadilan HAM menyebutkan bahwa genosida dipandang sebagai pelanggaran HAM yang berat. Tindakannya meliputi pembunuhan, menyebabkan penderitaan serius, pemusnahan, pemaksaan oleh kelompok tertentu, hingga pemindahan anak secara paksa dari suatu kelompok menuju kelompok lain. Dengan ketentuan tersebut, undang-undang secara tegas memberikan sanksi bagi setiap pelaku genosida [1].

Akhir-akhir ini, publik semakin sering memberikan pendapat mengenai isu genosida, misalnya genosida Israel terhadap Palestina. Berita mengenai Belakangan

ini, isu genosida banyak diangkat dalam berbagai portal berita daring seperti Kompas dan Detikcom. Namun, kedua portal tersebut tidak menyediakan API yang memungkinkan akses lebih luas terhadap berita-berita yang mereka publikasikan. Oleh karena itu, diperlukan sumber lain yang memuat banyak informasi terkait genosida sekaligus menyediakan API untuk mengakses data tersebut. Salah satu media sosial yang paling banyak digunakan masyarakat dan memiliki dukungan API adalah X. *Platform* ini berfungsi sebagai sarana bagi pengguna untuk berkomunikasi serta memperoleh informasi mengenai berbagai topik. Banyak pengguna yang menyampaikan pendapat mereka mengenai genosida melalui X, namun opini-opini tersebut sering kali tidak jelas apakah bernada positif atau negatif. Berdasarkan permasalahan tersebut, diperlukan analisis sentimen untuk mengidentifikasi kecenderungan opini masyarakat, apakah bersifat positif, negatif, atau netral.

Machine learning terus mengalami perkembangan pesat dan memainkan peran krusial di berbagai sektor industri, termasuk dalam bidang analisis sentiment. Salah satu fungsi utama dalam *machine learning* adalah melakukan klasifikasi, yaitu proses untuk memprediksi label kelas yang bersifat diskrit. Label kelas tersebut berasal dari sekumpulan kemungkinan yang telah ditentukan sebelumnya. Beragam algoritma digunakan untuk melakukan klasifikasi, di antaranya BERT dan *Naïve Bayes Classifier*.

BERT dikenal dengan efisiensi pelatihannya yang tinggi serta kemampuannya menghasilkan performa unggul dalam berbagai penelitian terkait klasifikasi. Di sisi lain, *Naïve Bayes Classifier* populer karena kemudahan dan kecepatannya dalam memprediksi kelas pada kumpulan data uji. Algoritma ini juga efektif dalam menangani prediksi multi-kelas. Ketika asumsi kemandirian antar fitur terpenuhi, *Naïve Bayes Classifier* sering kali mampu memberikan hasil yang lebih baik dibandingkan model lainnya

1.2 Perumusan Masalah

Rumusan masalah dalam penelitian ini asalah sebagai berikut:

1. Bagaimana cara menerapkan Algoritma BERT untuk melakukan analisis terhadap sentimen yang diberikan oleh masyarakat berdasarkan *post* di X mengenai genosida.
2. Melakukan perbandingan performa yang dihasilkan dari algoritma BERT dan *Naïve Bayes Classifier.*

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut :

1. Mengimplementasikan Algoritma BERT dan *Naïve Bayes Classifier* untuk melakukan analisis sentimen berdasarkan *post* yang disampaikan oleh masyarakat melalui media sosial X mengenai Genosida.
2. Melakukan evaluasi performa yang dihasilkan dari algoritma BERT dan *Naïve Bayes Classifier* dalam melakukan klasifikasi suatu data *post*.

1.4 Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut:

1. Penelitian ini diharapkan dapat menjadi refrensi tinjauan bagi mahasiswa yang membutuhkan pembahasan terkait analisis sentimen.
2. Untuk mengetahui keunggulan antara kedua algoritma yaitu BERT dan *Naïve Bayes Classifier* dalam melakukan analisis sentimen.

1.5 Batasan Masalah

Dalam penelitian ini dataset yang digunakan yakni *post* pengguna X mengenai genosida. Pada penelitian ini bahwa genosida yang dimaksud tentang penghentian terhadap genosida.

1.6 Sistematika Penulisan Skripsi

Sistematik penulisan skripsi ini terdiri dari lima bab sebagai berikut:

BAB 1: PENDAHULUAN

Pada bab ini menjelaskan secara umum mengenai latar belakang, rumusan masalah, tujuan dan manfaat penelitian, serta batasan dari penelitian yang dilakukan terkait sentimen masyarakat terhadap genosida pada media sosial X.

BAB II: TINJAUAN PUSTAKA

Pada bab ini berisi teori-teori yang mendasari penelitian ini seperti Genosida, X, Analisis Sentimen, *Natural Language Processing* (NLP), *Naive Bayes Classifier, Bidirectional Encoder Representations From Transformer* (BERT), T*ext Processing, Term Frequency Invers Document Frequency* (TF-IDF), *Classification Report, Confusion Matrix, Cross Industry Standard Process For Data Mining* (CRISP-DM), Google *Colaboratory*, Python, selain itu juga memuat penelitain terdahulu.

BAB III: METODOLOGI PENELITIAN

Pada bab ini membahas mengenai waktu dan tempat penelitian, serta tahapan penelitian yang dilakukan menggunakan *Cross-Industry Standard Process For Data Mining.*

BAB IV: HASIL DAN PEMBAHASAN

Pada bab ini, memuat pembahasan serta hasil yang diperoleh dari penelitian ini yang meliputi *Business Understanding, Data Understanding, Data Preparation (cleaning data, case folding, tokenization, word normalization, stopword removal, stemming, labelling* data dan pembobotan TF-IDF), *Modelling* menggunakan BERT dan *Naïve Bayes Classifier*, *Evaluation* menggunakan *classification report* dan *confusion matrix.*

BAB V: KESIMPULAN DAN SARAN

Pada bab ini memuat hasil penelitian yang sesuai dengan tujuan yaitu membandingkan metode BERT dan *Naïve Bayes Classifier* dalam menganalisis sentimen masyarakat terhadap berita genosida.

DAFTAR PUSTAKA

Pada bab ini memuat daftar literatur yang digunakan pada penelitian dan lampiran.

II. TINJAUAN PUSTAKA

2.1 Genosida

Istilah genosida berasal dari kata “geno” (bahasa Yunani: ras) dan “*cidium*” (bahasa latin: membunuh). Berdasarkan Statuta Roma dan Pasal 7 huruf UU Pengadilan HAM, genosida adalah perbuatan yang dijalankan dengan tujuan memusnahkan seluruh atau sebagian kelompok suatu bangsa, ras, etnis, atau agama melalui cara seperti pembunuhan, menyebabkan penderitaan fisik dan mental berat, menciptakan kondisi yang mengarah pada kehancuran kelompok, mencegah kelahiran, atau memindahkan anak-anak secara paksa ke kelompok lain [1]. Pasal 8 dalam Undang-Undang Nomor 26 Tahun 2000 tentang Pengadilan Hak Asasi Manusia mendefinisikan kejahatan genosida sebagai segala tindakan yang bertujuan untuk menghancurkan atau memusnahkan, baik secara keseluruhan maupun sebagian, suatu kelompok yang didasarkan pada bangsa, ras, etnis, atau agama. Pasal ini juga merinci lima bentuk tindakan yang termasuk dalam kejahatan genosida, yaitu:

1. Membunuh anggota suatu kelompok.
2. Menyebabkan penderitaan fisik atau mental yang berat pada anggota kelompok.
3. Menciptakan kondisi kehidupan yang dirancang untuk menghancurkan kelompok tersebut, baik secara keseluruhan maupun sebagian.
4. Melakukan tindakan pemaksaan dengan tujuan mencegah kelahiran dalam kelompok tersebut.
5. Memindahkan anak-anak secara paksa dari satu kelompok ke kelompok lain [2].

Tindakan genosida merupakan kejahatan luar biasa terhadap kemanusiaan yang melanggar nilai moral dan hak asasi manusia paling mendasar. Tindakan ini tidak hanya menyisakan luka yang mendalam bagi korban dan masyarakat, tetapi juga mengancam perdamaian dan stabilitas dunia. Oleh karena itu, segala bentuk tindakan genosida harus dikutuk dan dihentikan sepenuhnya agar tidak terulang kembali di masa depan.

2.2 *Platform* X

X sebelumnya dikenal sebagai Twitter hingga Juli 2023, X adalah *platform* media sosial dan layanan jejaring sosial yang dijalankan oleh X Corp, penerus Twitter, Inc. X merupakan *platform* media sosial yang banyak digunakan media *online* karena jangkauannya luas, komunikasi berlangsung cepat, serta berpotensi besar untuk periklanan. Dengan lebih dari 330 juta pengguna aktif bulanan, X mampu membantu penyebaran informasi secara cepat dan menjangkau audiens yang beragam [3]. Perubahan nama dan logo Twitter menjadi X merupakan bagian dari visi besar Elon Musk, pemilik baru *platform* ini, untuk menciptakan sebuah aplikasi serba bisa. Musk menyatakan bahwa transformasi ini bertujuan mendukung kebebasan berbicara dan mempercepat pengembangan X sebagai aplikasi multifungsi. Menurut Musk, nama Twitter awalnya cocok ketika *platform* tersebut hanya digunakan untuk berbagi 140 karakter layaknya kicauan burung. Namun, seiring perkembangan fitur yang kini memungkinkan unggahan video berdurasi panjang dan berbagai jenis konten lainnya, nama Twitter dianggap tidak lagi relevan. Oleh karena itu, Musk memutuskan untuk meninggalkan simbol burung yang telah lama menjadi ikon *platform* ini. Salah satu fitur utama yang direncanakan Musk adalah integrasi layanan keuangan secara menyeluruh di X. Ia berencana memasukkan dukungan untuk mata uang *Kripto Dogecoin*, yang secara terbuka didukungnya. Dengan fitur ini, pengguna X akan dapat melakukan transaksi seperti pembayaran, transfer dana, dan investasi langsung melalui aplikasi [4].

Seiring dengan transformasinya menjadi *platform* yang lebih luas, X juga menjadi ruang publik digital penting di tengah berbagai isu kemanusiaan global, termasuk genosida. Saat ini, masyarakat dari berbagai belahan dunia menggunakan X sebagai sarana untuk menyuarakan opini, menyebarkan kesadaran, dan membangun solidaritas terhadap korban genosida, Di tengah keterbatasan akses media arus utama, unggahan-unggahan di X memungkinkan siapa pun untuk menyampaikan sudut pandangnya, menyebarkan informasi yang mungkin tidak terangkat di media besar, hingga menyuarakan desakan untuk penghentian kekerasan dan genosida. Melalui fitur *quotes post*, *threads*, dan *repost*, pengguna dapat mendorong diskusi global, membagikan analisis, bahkan menyerukan tindakan nyata dari lembaga internasional. Banyak seruan dari masyarakat sipil seperti boikot produk pendukung rezim penindas, dukungan terhadap resolusi PBB, hingga tekanan terhadap para pemimpin dunia untuk menghentikan agresi militer. Dengan demikian, X kini bukan hanya *platform* sosial biasa, tapi juga arena perlawanan digital dan perjuangan opini dalam isu-isu kemanusiaan.

2.3 *Machine Learning*

Machine Learning atau pembelajaran mesin merupakan bagian dari kecerdasan buatan (*Artificial Intelligence*) yang bersifat *multidisipliner*, karena melibatkan berbagai bidang ilmu seperti statistika, probabilitas, ilmu komputer, teori informasi, psikologi, neurobiologi, dan filsafat. Tujuan utama *machine learning* adalah membuat sistem yang mampu belajar dari data, mengenali pola, dan mengambil keputusan secara otomatis tanpa perlu diprogram secara jelas. Perkembangan *machine learning* berawal dari penelitian tentang bagaimana komputer dapat meniru cara kerja otak manusia dalam belajar dan berpikir. Seiring waktu, bidang ini berkembang pesat mulai dari penemuan perceptron oleh Frank Rosenblatt yang menjadi dasar dari jaringan saraf tiruan, hingga munculnya berbagai Teknik *machine learning* seperti *Supervised learning, Unsupervised learning, dan Reinforcement Learning* [5].

1. *Supervised Learning*

Supervised learning merupakan metode pembelajaran dalam *machine learning* yang menggunakan data berlabel, yaitu data yang memiliki pasangan antara *input* (fitur) dan *output* (label) yang telah diketahui. Tujuan utama metode ini adalah membangun model prediktif yang mampu mempelajari hubungan antara *input* dan *output* untuk kemudian digunakan dalam memprediksi data baru. Proses pelatihannya dilakukan dengan cara algoritma membandingkan hasil prediksi dengan label sebenarnya, kemudian memperbaiki model berdasarkan kesalahan yang ditemukan. *Supervised learning* banyak diterapkan untuk memprediksi kejadian di masa depan, seperti mendeteksi transaksi penipuan kartu kredit, menentukan nasabah yang berpotensi mengajukan klaim asuransi, mengklasifikasikan jenis bunga iris, hingga sistem rekomendasi film. Metode ini terbagi menjadi dua jenis, yaitu:

a. Klasifikasi (*Classification*) label bersifat diskret, misalnya menentukan sentimen positif, negatif, atau netral.
b. Regresi (*Regression*) label bersifat kontinu, seperti memprediksi nilai harga, suhu, atau usia.

Dengan demikian, *supervised learning* berfokus pada pembelajaran dari data berlabel untuk menghasilkan model yang dapat memprediksi hasil dari data baru secara akurat.

2. *Unsupervised Learning*

Unsupervised learning merupakan metode pembelajaran mesin yang bekerja pada data tanpa label. Tujuannya adalah untuk menemukan pola, hubungan, atau struktur tersembunyi di dalam data tanpa mengetahui hasil yang benar sebelumnya. Metode ini sering digunakan untuk mengelompokkan data (*clustering*) seperti segmentasi pelanggan berdasarkan kesamaan atribut, atau untuk mengidentifikasi pola penting dan data yang menyimpang (*outlier*). Beberapa algoritma yang umum digunakan antara lain *K-Means Clustering, Self-Organizing Maps, Nearest Neighbor Mapping,* dan *Singular Value Decomposition* (SVD). Dalam prosesnya, algoritma

unsupervised learning membangun model yang menyesuaikan parameter secara otomatis untuk meringkas keteraturan yang terdapat pada data.

3. ***Reinforcement Learning***

Metode ini sering digunakan dalam bidang robotika, permainan (*gaming*), dan navigasi. *Reinforcement learning* merupakan teknik pembelajaran di mana sistem berinteraksi dengan lingkungan yang dinamis untuk mencapai suatu tujuan tertentu tanpa adanya "guru" yang secara eksplisit memberi tahu apakah tindakannya sudah mendekati tujuan atau belum. Dalam *reinforcement learning*, algoritma belajar melalui proses coba-coba (*trial* and *error*) untuk menemukan tindakan mana yang memberikan imbalan (*reward*) paling besar Sebagai contoh, pada permainan catur, sistem *reinforcement learning* belajar bermain dengan melawan lawan mainnya, mencoba berbagai strategi hingga menemukan cara terbaik untuk menang. Jenis pembelajaran ini memiliki tiga komponen utama, yaitu:

a. *Learner* (agen/pembelajar) pihak yang belajar dan mengambil keputusan,
b. *Environment* (lingkungan) kondisi atau situasi tempat agen berinteraksi,
c. *Actions* (aksi) tindakan yang diambil agen untuk memengaruhi lingkungan.

Tujuan utamanya adalah agar agen dapat memilih tindakan yang memaksimalkan imbalan (*reward*) yang diharapkan dalam jangka waktu tertentu. Agen akan mencapai tujuan lebih cepat jika mengikuti kebijakan (*policy*) yang baik. Dengan demikian, tujuan utama *reinforcement learning* adalah mempelajari kebijakan terbaik untuk memperoleh hasil optimal.

Pada era modern, kemajuan teknologi komputasi dan ketersediaan data dalam jumlah besar membuat *machine learning* semakin penting. Teknik ini digunakan untuk menganalisis data kompleks, memprediksi tren, dan menghasilkan keputusan cerdas secara *real-time*. Model *machine learning* mampu meningkatkan efisiensi dan akurasi dalam berbagai bidang seperti keuangan, kesehatan, dan analisis teks. *Machine learning* terus berkembang seiring dengan munculnya algoritma baru yang lebih efisien dan adaptif. Pengembangan algoritma ini bertujuan untuk menciptakan sistem yang lebih cerdas, tangguh, dan mudah digunakan di dunia nyata.

2.4 *Natural Language Processing* (NLP)

Pemrosesan Bahasa Alami *Natural Language Processing* (NLP) adalah bidang teknik yang menggunakan komputasi dan kecerdasan buatan untuk menganalisis dan merepresentasikan bahasa manusia. Tujuannya adalah untuk membangun interaksi yang lebih baik antara manusia dan computer [6]. Secara umum, analisis bahasa dalam NLP dilakukan melalui beberapa teknik, termasuk sintaksis, semantik, dan *pragmatic* [7]. Dalam praktiknya, NLP diterapkan dalam berbagai bidang terkait pemrosesan bahasa alami, antara lain [8]:

a. *Speech Recognition*, digunakan untuk mengubah ucapan manusia menjadi teks yang dapat dipahami oleh komputer.
b. *Part-of-speech tagging*, berfungsi mengidentifikasi jenis kata dalam teks, seperti kata benda, kata kerja, atau kata sifat.
c. *Named Entity Recognition*, digunakan untuk menemukan dan mengenali entitas tertentu dalam teks, seperti nama orang, tempat, atau tanggal.
d. *Machine Translation*, diterapkan untuk menerjemahkan teks dari satu bahasa ke bahasa lain secara otomatis.
e. *Question Answering*, digunakan untuk memahami pertanyaan berbentuk teks dan menemukan jawaban yang sesuai dari teks tersebut.
f. *Text Classification*, berfungsi mengelompokkan teks ke dalam kategori tertentu berdasarkan isinya. Salah satu contoh penerapan *text classification* yang populer adalah analisis sentimen, yang bertujuan untuk mengidentifikasi sentimen atau emosi yang terkandung dalam teks. Dalam analisis sentimen, teks seperti ulasan produk, ulasan film, atau posting di media sosial akan dikategorikan sesuai dengan sentimen yang diungkapkannya. Biasanya metode yang populer digunakan pada *text classification* yaitu Metode *Naïve Bayes Classifier* kemudian ada beberapa penelitian juga yang melakukan penelitian dengan menggunakan metode BERT.

2.5 Analisis Sentimen

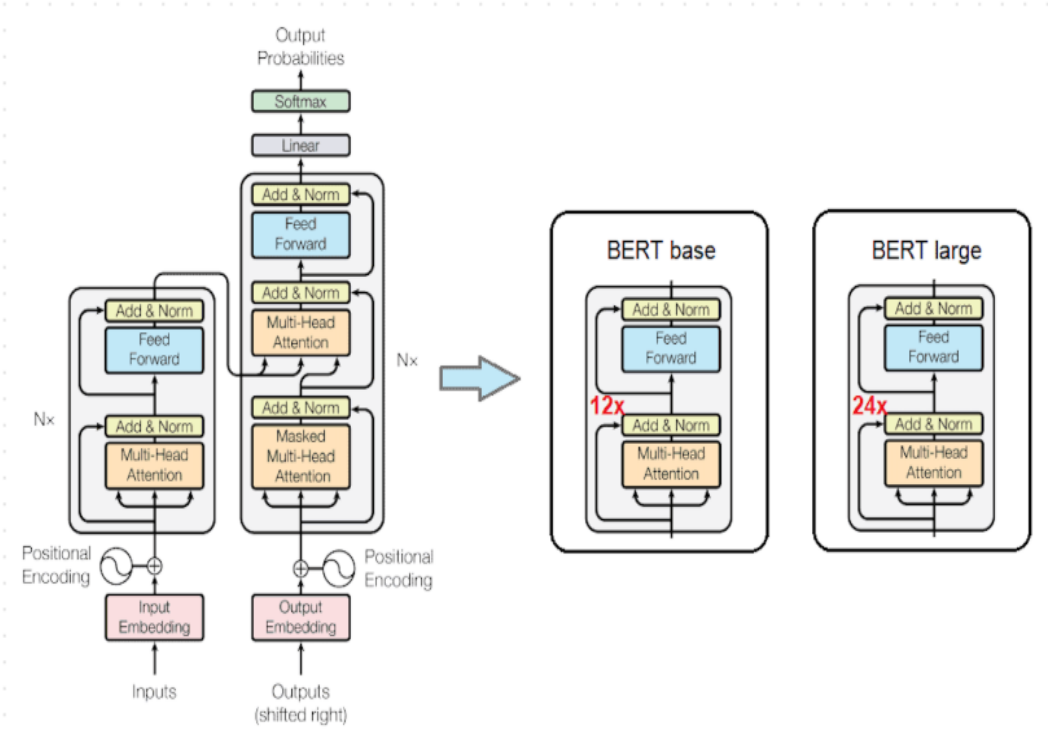
Analisis sentimen adalah bagian dari *Natural Language Processing* (NLP) yang berfungsi untuk mengenali serta mengambil informasi subjektif berupa opini atau

sentimen dari suatu teks. Analisis sentimen atau yang dapat disebut juga *opinion mining* adalah bidang studi yang dapat menganalisis sentimen, evaluasi, opini, penilaian, sikap, individu, isu, peristiwa, tema, dan atributnya. Penggunaan analisis sentimen seringkali dimanfaatkan untuk menghasilkan tanggapan terhadap layanan jasa, produk, atau topik tertentu, dengan tujuan untuk mendapatkan evaluasi kedepannya [9]. Tugas utama dari analisis sentimen yaitu untuk mengelompokkan teks yang terdapat pada suatu dokumen atau kalimat, kemudian akan ditentukan megenai pendapat yang terkandung dalam dokumen ataupun kalimat tersebut, apakah tergolong sentimen positif, sentimen netral, atau, sentimen negatif. Terdapat dua pendekatan yang bisa digunakan dalam analisis sentimen, yaitu pendekatan berbasis leksikon (*lexicon-based*) dan pendekatan berbasis *machine learning*. Pendekatan *lexicon-based* bergantung pada ketersediaan atau penyusunan sebelumnya terhadap kamus leksikal yang sesuai, sedangkan metode *machine learning* melakukan klasifikasi teks secara otomatis. namun umumya memerlukan data latih (data *training sets*) yang bersumber dari pemrograman manusia [10].

2.6 *Bidirectional Encoder Representations from Transformers* (BERT)

Bidirectional Encoder Representations from Transformers (BERT) adalah algoritma *deep learning* yang dirancang khusus untuk mengolah *Natural Language Processing* (NLP), sehingga mampu memahami hubungan antar kata dalam suatu kalimat. BERT memiliki arsitektur *truly bidirectional* yang memungkinkan model untuk memahami konteks secara simultan dari kiri ke kanan dan dari kanan ke kiri, dengan memanfaatkan jaringan yang sama. Algoritma ini menggunakan *transformer*, yaitu mekanisme yang mempelajari hubungan kontekstual antar kata dalam teks melalui mekanisme *self-attention*. Secara umum, *transformer* terdiri dari dua komponen utama, yakni *encoder* dan *decoder*. *Encoder* berfungsi untuk memproses teks *input*, sedangkan *decoder* bertugas membuat prediksi untuk tugas tertentu. Namun, dalam BERT, hanya *encoder* yang diperlukan karena tujuannya adalah membangun model bahasa. BERT terdiri dari dua tahapan utama yaitu *pre-training* dan *fine-tuning*. Pada tahap *pre-training*, model dilatih menggunakan data berskala besar yang tidak berlabel untuk menyelesaikan berbagai tugas. Setelah itu,

pada tahap *fine-tuning*, model diinisiasi dengan parameter dari *pre-trained* model dan kemudian diadaptasi dengan data berlabel untuk menyelesaikan tugas spesifik (*downstream tasks*). Meskipun model *fine-tuning* untuk setiap tugas berbeda, semua dimulai dari parameter *pre-trained* yang sama. BERT mampu menghasilkan performa yang sangat baik pada beragam tugas NLP, seperti *question answering*, *natural language inference*, klasifikasi teks, dan evaluasi pemahaman bahasa secara umum. Model ini memiliki dua jenis utama, yaitu BERT-*base* dan BERT-*large*. BERT-*base* terdiri dari 12 *layer*, ukuran *hidden* 768, memiliki 12 *attention heads*, dan sekitar 110 juta parameter. Sementara itu, BERT-*large* lebih besar dengan 24 *layer*, *hidden* size 1024, 16 *attention heads*, dan total sekitar 340 juta parameter [11].



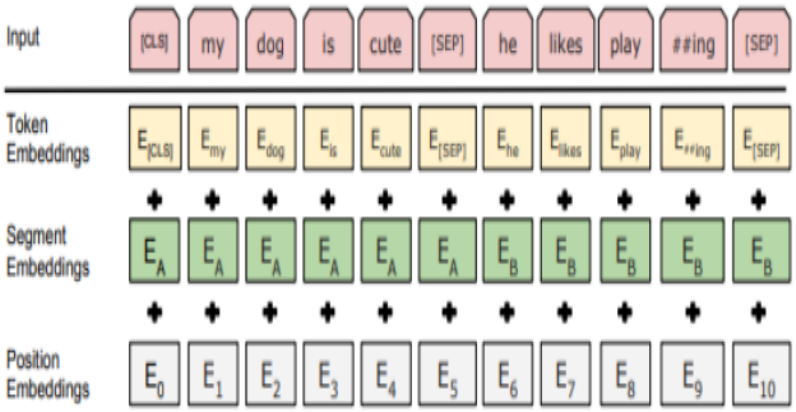
Gambar 2.1 Arsitektur BERT [12].

Berdasarkan gambar 2.1 BERT dibangun menggunakan arsitektur *deep neural network* bernama *Transformer*. Model ini menerima *input* dalam bentuk vektor numerik yang diperoleh melalui proses *word embedding*. Setiap token pada urutan *input* diubah menjadi representasi vektor melalui proses *embedding* tersebut. Karena *Transformer* tidak memiliki mekanisme berulang seperti RNN, informasi

posisi token harus ditambahkan secara eksplisit dengan menggunakan *positional encoding* yang digabungkan ke dalam *embedding* awal. Setelah itu, input yang telah diberi *positional encoding* diproses melalui mekanisme *multi-head self-attention.* Mekanisme ini memungkinkan model untuk memfokuskan perhatian pada berbagai bagian dalam urutan input di setiap lapisan sehingga mampu menangkap hubungan serta ketergantungan jarak jauh antar token. Hasil dari *self-attention* kemudian diteruskan ke jaringan *feed-forward* yang menerapkan transformasi non-linear pada setiap token secara mandiri. Agar proses pelatihan lebih stabil dan konvergen lebih cepat, digunakan residual *connections* dan *layer normalization*. Residual *connections* membantu aliran gradien tetap lancar, sedangkan *normalization* menjaga stabilitas distribusi nilai pada setiap lapisan. Secara keseluruhan, arsitektur *Transformer* terdiri dari dua komponen utama, yaitu *stack encoder* dan *stack decoder*. *Encoder* bertugas mengubah urutan input menjadi representasi terenkode, sementara *decoder* digunakan untuk menghasilkan urutan *output*. Pada bagian *decoder*, *self-attention* dibuat dalam bentuk *masked attention* sehingga setiap posisi hanya dapat memperhatikan token sebelumnya dan token saat ini. Mekanisme ini mencegah model melihat token-token masa depan dan memastikan proses generasi berjalan secara berurutan. Selain itu, selama proses *decoding*, *decoder* juga memanfaatkan informasi dari *encoder* untuk menghasilkan prediksi yang lebih akurat. Keluaran akhir dari *decoder* kemudian diproyeksikan ke dalam vektor probabilitas berdasarkan kosakata yang ada. Distribusi probabilitas inilah yang digunakan untuk memproduksi token output secara berurutan [13].

BERT mampu mempelajari hubungan kontekstual antar kata dalam sebuah kalimat melalui proses pelatihan pada kumpulan data teks berukuran besar. Model ini dilatih menggunakan dua jenis tugas utama, yaitu *Masked Language Modeling* (MLM) dan *Next Sentence Prediction* (NSP). Pada tugas MLM, sebagian kata dalam kalimat digantikan secara acak dengan token [MASK], kemudian model diminta menebak kata aslinya. Proses ini membantu BERT memahami makna kata berdasarkan konteks sekitarnya. Sementara itu, pada tugas NSP, model menerima dua kalimat dan harus menentukan apakah kalimat kedua merupakan kelanjutan yang logis dari kalimat pertama.

Tugas ini memungkinkan BERT mempelajari keterkaitan dan urutan antar kalimat [11]. Berikut merupakan representasi input yang digunakan BERT dalam pemrosesan:



Gambar 2.2 Representasi Input BERT [11].

1. Tokenisasi : Membagi teks menjadi token-token berupa kata-kata. BERT menggunakan tokenisasi *WordPiece*, yang berarti beberapa token dapat dibagi lagi menjadi sub-token.
2. Token *Embeddings* (penambahan token *special*) : BERT menambahkan dua token khusus ke awal dan akhir setiap kalimat, yaitu [CLS], [SEP], [PAD]. Token [CLS] Adalah *classification* token yang digunakan untuk merepresentasikan kalimat secara keseluruhan yang berada di awal kalimat, sedangkan token [SEP] adalah separator token yang biasanya akan berada di akhir kalimat yang digunakan untuk memisahkan kalimat dalam input yang berbeda dari urutan input dan token [PAD] adalah padding token yang biasanya akan digunakan untuk menyamakan input.

$$\mathbf{T'=\{[CLS],t1,t2,...,tm,[SEP],[PAD],...,[PAD]\}} \quad (1)$$

Keterangan:

T′ = urutan token setelah ditambah token khusus.

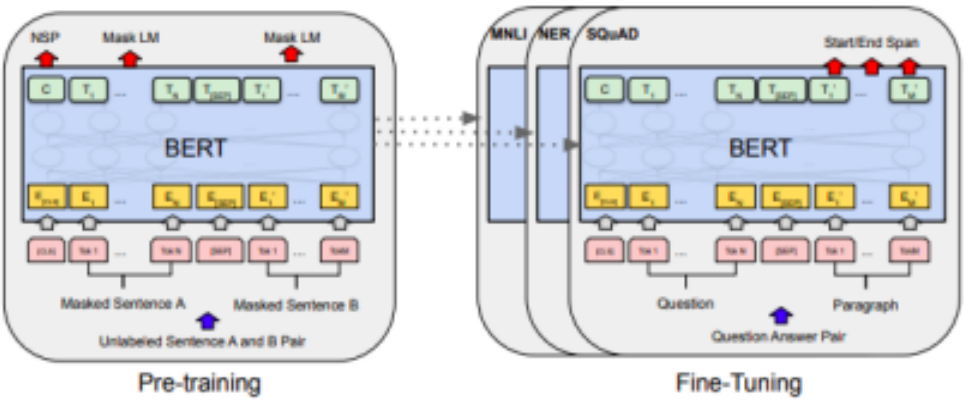
[CLS] = token klasifikasi (selalu di depan).

[SEP] = token pemisah/akhir kalimat.

[PAD] = padding, untuk menyamakan panjang sequence dengan maksimal panjang (L) yang ditentukan.

3. Konversi Token menjadi ID : Setiap token dalam input kemudian dikonversi menjadi ID token yang sesuai menggunakan kamus token yang telah ditetapkan. Selanjutnya, setiap ID token dikonversi menjadi vektor dengan mengambil nilai *embedding* dari matriks *embedding* kata yang telah dilatih sebelumnya. Matriks *embedding* menggambarkan setiap kata dalam ruang vektor yang terdiri dari banyak dimensi.
4. *Segment Embeddings* : Jika input terdiri dari dua kalimat, setiap token dalam *input* harus ditandai sebagai milik kalimat pertama atau kedua. Ini dilakukan dengan memberikan segmen ID 0 atau 1 ke setiap token, tergantung pada kalimat mana yang mengandung token tersebut.
5. *Position Embedding* : BERT memanfaatkan *position embedding* untuk menyisipkan informasi posisi absolut ke dalam representasi setiap token. Hal ini dilakukan dengan menambahkan vektor posisi yang telah ditetapkan sebelumnya pada vektor masing-masing token.

BERT memiliki dua paradigma pelatihan yaitu *pre-training* dan *fine tuning* yang ditunjukkan pada Gambar 2.3.



Gambar 2.3 *Pre-training* dan *Fine-tuning* pada BERT [11].

Pada gambar 2.3 merupakan *pre-training* dan *fine-tuning* pada BERT, berikut pengertiannya.

1. Pre-training

Pada tahap *pre-training*, BERT dilatih sebagai *unsupervised learning* karena menggunakan data tanpa label untuk mengenali pola dalam teks. Model ini dilatih pada *BooksCorpus* (800 juta kata) dan English Wikipedia (2,5 miliar kata). *Pre-training* BERT mencakup dua tugas utama, yaitu *Masked Language Modeling* (MLM) dan *Next Sentence Prediction* (NSP). MLM dilakukan dengan mengganti sebagian kata dalam kalimat menggunakan token [MASK] sehingga model harus menebak kata aslinya. Hal ini membantu BERT memahami arti kata berdasarkan konteks. Sementara itu, pada NSP, model menentukan apakah suatu kalimat merupakan lanjutan yang sesuai dari kalimat sebelumnya, sehingga melatih pemahaman hubungan antarkalimat.

2. Fine-Tuning

Fine-tuning adalah proses melatih model untuk tugas tertentu dengan menggunakan model yang sebelumnya sudah dilatih menggunakan dataset besar. Proses *fine-tuning* menjadi lebih mudah karena mekanisme *self-attention* pada *Transformer* memungkinkan BERT untuk menangani berbagai tugas hilir, baik yang melibatkan teks tunggal maupun pasangan teks. *Fine-tuning* ini sangat penting untuk model bahasa *pre-trained* seperti BERT, karena dengan menggunakan dataset besar, model BERT dapat mengerti konteks dengan lebih baik. Tanpa *pre-trained language* model, proses pelatihan akan membutuhkan lebih banyak sumber daya dan waktu. Dengan *pre-trained* model, tugas-tugas spesifik menjadi lebih sederhana karena model sudah dilatih menggunakan data yang besar, sehingga hanya perlu disesuaikan melalui *fine-tuning*. Pada *fine tuning* ini akan menggunakkan parameter *Batch Size* dan *Epoch*. *Batch size* merupakan banyaknya sampel data yang diproses sekaligus dalam satu iterasi melalui jaringan saraf. Nilai *batch size* menentukan berapa jumlah data yang diolah sebelum model melakukan pembaruan terhadap parameter-parameternya [14]. Ada beberapa hal yang harus diperhatikan untuk memilih ukuran pada *batch size* yaitu:

- Batasan Memori

Batch size kecil (misalnya 16, 32, dan seterusnya) hanya memerlukan sedikit memori, sehingga lebih sesuai digunakan pada perangkat dengan kapasitas *hardware* terbatas. *Batch size* berukuran besar (seperti 256, 512, dan lainnya) membutuhkan memori yang jauh lebih besar, namun dapat mempercepat proses training apabila tersedia GPU atau TPU dengan spesifikasi tinggi.

- Stabilitas Pelatihan

Batch size kecil menghasilkan gradien yang lebih berisik. Hal ini bisa membantu model keluar dari titik minimum lokal, tetapi juga dapat membuat proses pelatihan menjadi kurang stabil. Sedangkan *batch size* besar memberikan gradien yang lebih halus dan stabil serta memungkinkan konvergensi lebih cepat. Namun, model bisa berakhir pada minimum yang "tajam", sehingga performanya kurang baik ketika diuji pada data baru.

- Kecepatan Pelatihan

Batch kecil biasanya membuat pelatihan lebih lambat karena model harus melakukan lebih banyak langkah pembaruan untuk menyelesaikan satu *epoch*. Sedangkan *batch size* besar dapat mempercepat *training* dengan mengurangi jumlah pembaruan yang diperlukan dalam setiap *epoch*.

Sedangkan *epoch* menunjukkan berapa kali keseluruhan dataset pelatihan diproses oleh model. Misalnya, jika terdapat 1000 data latih dan Anda menetapkan 10 *epoch*, berarti model akan mempelajari seluruh dataset tersebut sebanyak 10 kali. Nilai *epoch* menentukan seberapa sering model dilatih menggunakan seluruh data. Menentukan jumlah *epoch* yang tepat sangat penting agar model mampu menghasilkan performa optimal tanpa mengalami *overfitting* [15].

2.7 *Naïve Bayes Classifier*

Salah satu metode klasifikasi yang banyak digunakan dalam *text mining* adalah algoritma *Naïve Bayes Classifier*. Algoritma ini efektif dalam memproses data berukuran besar serta mampu menghasilkan akurasi yang cukup tinggi [16]. *Naïve Bayes* merupakan metode klasifikasi yang didasarkan pada prinsip probabilitas dan

statistik, yang pertama kali dikembangkan oleh ilmuwan Inggris, Thomas Bayes. Metode ini bertujuan untuk memprediksi kemungkinan suatu kejadian di masa depan berdasarkan data historis yang ada, konsep ini dikenal sebagai Teorema Bayes. Kinerja sistem yang menggunakan algoritma *Naïve Bayes Classifier* sangat bergantung pada data yang tersedia dan digunakan sebagai data latih. Jika data latih dapat mewakili sebagian besar atau keseluruhan data yang dimiliki, maka sistem klasifikasi akan memiliki performa yang baik. Jika kinerja sistem klasifikasi tersebut memadai, sistem tersebut dapat diterapkan untuk mengklasifikasikan data dalam jumlah yang lebih besar [17].

Pada proses klasifikasi teks dengan *Naïve Bayes Classifier*, terdapat dua tahap utama, yaitu tahap pelatihan (*training*) dan pengujian (*testing*). Pada tahap pertama, dilakukan pelatihan menggunakan data sentimen yang kelasnya sudah diketahui untuk membangun model probabilistik. Selanjutnya, pada tahap kedua, dilakukan klasifikasi sentimen pada data yang kelasnya belum diketahui. Secara umum, rumus Teorema Bayes dapat dituliskan sebagai berikut [18].

$$\frac{P(H|X)=P(X|H)\ P(H)}{P(X)} \qquad (2)$$

Keterangan :

X = Data dengan *class* yang belum diketahui

H = Hipotesis data X merupakan suatu *class* spesifik

$P(H|X)$ = Probabilitas hipotesis H berdasarkan kondisi x (posteriori prob)

P(H) = Probabilitas hipotesis H (prior prob)

P(X|H) = Probabilitas X berdasarkan kondisi tersebut

B P(X) = Probabilitas dari X

Berdasarkan fungsinya algoritma *naïve bayes classifier* digolongkan menjadi 3 tipe antara lain *Bernoulli Naïve Bayes* jenis *Naïve Bayes* yang diterapkan pada data kategorikal dengan dua kemungkinan hasil untuk setiap atribut fitur. Jenis ini menggunakan nilai biner dalam proses klasifikasinya. *Bernoulli Naïve Bayes* bekerja dengan data diskrit dan hanya menerima fitur dalam bentuk nilai biner seperti benar atau salah, ya atau tidak, berhasil atau gagal, 0 atau 1, dan seterusnya.

Kemudian *Gaussian Naïve Bayes* jenis *Naïve Bayes* yang digunakan untuk menghitung probabilitas data kontinu yang berkaitan dengan setiap fitur numerik dalam kaitannya dengan kelas tertentu. *Gaussian Naïve Bayes* ditentukan oleh dua parameter, yaitu rata-rata dan standar deviasi. Pada penelitian ini menggunakan *Multinomial Naïve Bayes* dikarenakan pada jenis ini sering digunakan untuk menyelesaikan masalah klasifikasi dokumen yang panjang dengan jumlah kosakata yang besar. Algoritma ini mengasumsikan independensi antara kemunculan kata-kata dalam dokumen, tanpa memperhitungkan urutan dan konteks kata tersebut. Selain itu, algoritma ini juga mempertimbangkan frekuensi kemunculan kata dalam dokumen. Probabilitas bahwa suatu dokumen d berada dalam kelas c dapat dihitung menggunakan persamaan berikut:

$$P(c|d) \propto P(c) \prod_k^n = P(tk|c) \qquad (3)$$

Keterangan:

$P(c|d)$ = Probabilitas dokumen d berada di kelas c

$P(c)$ = Prior probabilitas suatu dokumen berada di kelas c

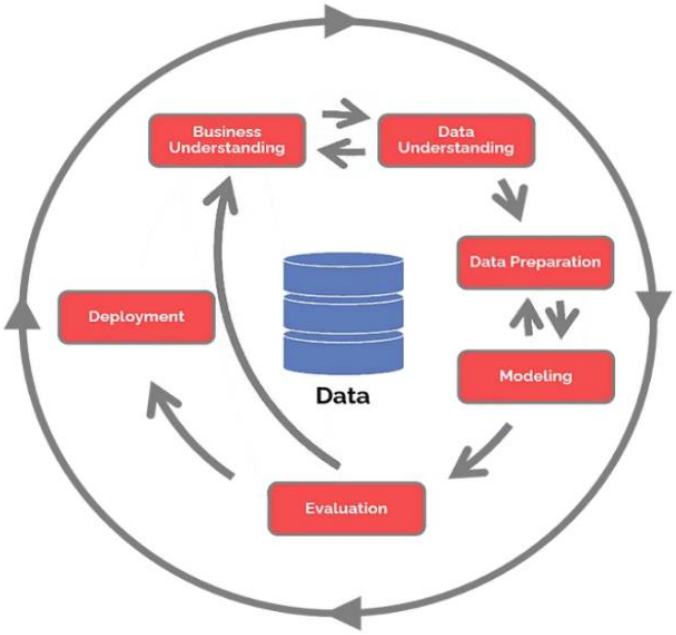
$\{t1,t1,t1, \dots ,tn\}$ = Token dalam dokumen d yang merupakan bagian dari *vocabulary* dengan jumlah n.

$P(tk|c)$ = Probabilitas bersyarat *term* tk berada di dokumen pada kelas c

2.8 *Cross Industry Standard Process ForData Mining* (CRISP-DM)

CRISP-DM (*Cross Industry Standard Process for Data Mining*) adalah metodologi yang banyak digunakan dalam *text mining*. Metode ini umum diterapkan untuk menyelesaikan berbagai masalah bisnis yang berkaitan dengan data mining. CRISP-DM pertama kali diperkenalkan dan dikembangkan pada tahun 1996 oleh beberapa analis industri, di antaranya dari NCR, SPSS, dan *Daimler Chrysler*. CRISP-DM tidak dirancang melalui pendekatan teoretis-akademis yang berlandaskan prinsip teknis, maupun melalui keputusan tertutup dari kelompok ahli

tertentu. Pendekatan semacam itu memang pernah digunakan sebelumnya dalam penyusunan metodologi, tetapi jarang menghasilkan standar yang praktis, efektif, dan dapat diterima secara luas. Keberhasilan CRISP-DM justru terletak pada fondasinya yang kuat, yakni pengalaman nyata para praktisi dalam melaksanakan proyek data mining. Oleh karena itu, kontribusi gagasan dan usaha dari banyak praktisi menjadi faktor penting dalam terwujudnya CRISP-DM. Metodologi ini terdiri dari enam tahap: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, *dan Deployment* [19]. Namun pada penelitian ini hanya sampai pada tahap evaluasi.



Gambar 2.4 Tahapan Metodologi CRISP-DM

2.8.1 *Business Understanding* (Pemahaman Bisnis)

Pemahaman bisnis merupakan tahap awal dalam proses CRISP-DM. Pada tahap ini, fokus utama adalah memahami tujuan dan kebutuhan dari perspektif bisnis, yang kemudian diterjemahkan ke dalam permasalahan di bidang penambangan data. Selanjutnya, dibuatlah rencana yang jelas untuk mencapai tujuan tersebut

2.8.2 Data *Understanding* (Pemahaman Data)

Tahap pemahaman data adalah proses pengumpulan data awal yang relevan dengan tujuan bisnis yang telah ditetapkan sebelumnya. Setelah itu, dilakukan analisis terhadap data tersebut untuk memahami jenis data yang akan digunakan dan mengevaluasi kualitasnya. Tahap ini juga memungkinkan untuk kembali ke tahap sebelumnya, yaitu pemahaman bisnis, untuk memastikan bahwa data yang dikumpulkan dapat mendukung pencapaian tujuan yang telah ditentukan.

2.8.3 Data *Preparation* (Persiapan Data)

Tahap persiapan data adalah proses yang dilakukan untuk menghasilkan dataset akhir dari data mentah melalui langkah-langkah *pre-processing*, sebelum melanjutkan ke tahap pemodelan. *Pre-processing* adalah tahapan yang bertujuan untuk mengolah data mentah menjadi format yang lebih mudah dipahami, sehingga dapat digunakan dalam pemodelan. Langkah-langkah dalam *pre-processing* meliputi pembersihan data, *case folding*, tokenisasi, dan penghapusan kata yang tidak berarti (*stopword removal*). Tahap persiapan data ini dapat dilakukan secara berulang sesuai dengan kebutuhan yang diinginkan. Kemudian dilakukan *Text pre-processing* adalah tahap awal dalam mempersiapkan teks agar dapat diolah lebih lanjut menjadi data yang dapat digunakan. Karena teks tidak bisa langsung diproses oleh algoritma, proses ini diperlukan untuk mengubahnya menjadi data numerik. Proses ini melibatkan beberapa langkah pembersihan dokumen [20]. Dalam penelitian ini, *pre-processing* mencakup:

a. *Tokenizing*, yaitu memecah deskripsi yang awalnya berupa kalimat menjadi kata-kata.
b. *Cleaning* Data, membersihkan data dari karakter-karakter yang tidak diperlukan seperti tautan, nama pengguna, angka, simbol, dan kata-kata yang mengganggu proses analisis.
c. *Stemming*, mengubah kata-kata dalam bahasa Indonesia menjadi bentuk dasarnya dengan menghilangkan imbuhan awalan, akhiran, dan sisipan.
d. *Stopword Removal*, menghapus kata-kata yang sering muncul tetapi tidak relevan, seperti kata hubung yang tidak berpengaruh pada proses klasifikasi.

e. *Case Folding*, mengubah teks menjadi huruf kecil atau huruf besar sesuai kebutuhan klasifikasi.
f. *Word Normalization*, mengoreksi kata-kata yang tidak tepat dengan menggunakan kamus agar memiliki arti yang sesuai, sehingga tidak menambah dimensi vektor yang bisa memperlambat proses komputasi [21].

Pada Penelitian ini kamus yang digunakan yaitu kamus *colloquial* Indonesian *lexicon. Colloquial Indonesian Lexicon* ragam bahasa variasi bahasa menurut pemakaiannya, yang berbagai macam dari topik yang dibicarakan, menurut dari; teman bicara, orang yang dibicarakan, dan bagi media pembicaranya. Ragam bahasa sering menjadi pembeda dalam kreativitas, ekspresi komunikasi, serta membedakan individu atau kelompok dalam masyarakat. Dalam ragam bahasa gaul merupakan bentuk variasi bahasa yang selalu berkembang seiring zaman, terutama pada kalangan remaja. Media sosial menjadi ruang interaksi komunikasi bagi remaja sehingga bahasa gaul cepat menyebar danmembentuk kosakata baru, seperti pada aplikasi X atau dulu bernama Twitter. Media sosial X merupakan salah satu aplikasi yang berpengaruh terhadap perkembangan bahasa. Kehidupan masyarakat yang semakin modern menjadikan hal ini bagian yang tak dapat dipisahkan, termasuk remaja. Melalui media sosial, seseorang dapat dengan mudah mencari berbagai informasi dan budaya. Pengguna media sosial sering mempelajari suatu bahasa, terlebih bahasa gaul melalui berbagai cara, seperti interaksi komunikasi dengan pengguna lain, melalui tren yang viral, serta konten lucu atau meme [22].

2.8.4 *Modelling* (Pemodelan)

Tahap pemodelan bertujuan untuk merancang model yang akurat guna menganalisis kemungkinan dari data yang masuk dan mengkategorianya berdasarkan perhitungan peluang yang telah dilakukan sebelumnya. Sebelum algoritma pemodelan diterapkan, dataset terlebih dahulu dibagi menjadi dua bagian, yaitu data *training* dan data *testing*. Data *training* digunakan untuk membentuk *classifier* atau memberikan label pada data yang akan dipelajari oleh model. Sementara itu, data *testing* berfungsi untuk menilai kinerja sistem yang telah dilatih.

2.8.5 *Evaluation* (Pengujian)

Tahap ini dilakukan untuk menilai kualitas model yang telah dibuat. Hasil evaluasi tersebut menjadi dasar untuk menentukan apakah model sudah memenuhi kriteria yang diinginkan. Jika performanya belum optimal, maka proses pengembangan model dapat diulang kembali.

2.9 *Term Frequency Invers Document Frequency* (TF-IDF)

TF-IDF adalah teknik yang digunakan untuk memberikan bobot pada hubungan antara sebuah kata dengan dokumen. Metode ini bekerja dengan memberikan bobot yang lebih tinggi pada istilah yang sering muncul dalam dokumen tertentu, tetapi jarang muncul di banyak dokumen dalam satu korpus. TF-IDF mengombinasikan dua konsep dalam menghitung bobot, yaitu frekuensi kemunculan kata dalam dokumen tertentu (TF) dan kebalikan dari frekuensi dokumen yang mengandung kata tersebut (IDF). TF (*Term Frequency*) mengukur seberapa sering kata muncul dalam sebuah dokumen, di mana semakin sering kata muncul, semakin tinggi bobotnya. Sementara itu, IDF (*Inverse Document Frequency*) mengurangi dominasi istilah yang sering muncul di banyak dokumen, karena istilah yang terlalu umum dianggap kurang penting. Sebaliknya, kata yang jarang muncul di dokumen lebih diperhatikan dan dianggap lebih penting. Pembobotan akan memperhitungkan faktor kebalikan frekuensi dokumen yang mengandung suatu kata (*Inverse Document Frequency*). Adapun rumus umum untuk pembobotan TF-IDF ini adalah sebagai berikut [23].

$$idf_t = log \frac{N}{df_t} \quad (4)$$

$$f - idf_{d,t} = tf_{d,t}\, x\, idf_{d,t} \quad (5)$$

Dimana:

tf = banyaknya kata yang dicari pada sebuah dokumen.

idf_t = *Inversed Dokumen Frequency*.

N = Total Dokumen

df_t = Jumlah dokumen yang mengandung *term* t

d = Dokumen ke-d

t = kata ke-t dari kata kunci.

$tf - idf_{d,t}$ = bobot dokumen ke-d terhadap kata ke-t

2.10 *Classification Report*

Classification report adalah salah satu metode untuk mengevaluasi kinerja model dalam *machine learning*. Laporan ini menampilkan metrik seperti akurasi (*accuracy*), presisi (*precision*), *recall*, *F1-score*, dan *support* berdasarkan model yang telah dibangun. Dengan menggunakan *classification report*, kita dapat memahami kinerja keseluruhan dari model yang telah dilatih dengan lebih baik. *Classification report* mencakup beberapa indikator metrik yang mengukur performa model, seperti akurasi, presisi, *recall*, *F1-score*, dan *support*, seperti yang ditunjukkan pada gambar 2.5 di bawah ini [24].

	precision	recall	f1-score	support
ham	0.99	0.99	0.99	1587
spam	0.93	0.92	0.92	252
accuracy			0.98	1839
macro avg	0.96	0.95	0.96	1839
weighted avg	0.98	0.98	0.98	1839

Gambar 2.5 *Classification Report*

Berikut ini merupakan penjelasan metriks evaluasi dari *classification report*:

1. Akurasi (Accuracy)

   Akurasi (*accuracy*) merupakan perbandingan prediksi bernilai benar (*True Positive* dan *True Negative)* dengan keseluruhan data yang ada. Metrik ini memberikan informasi tentang seberapa baik model dapat melakukan klasifikasi yang tepat. Nilai *accuracy* memberikan perbandingan antara pengamatan yang diklasifikasikan dengan benar (positif dan negatif) dengan

jumlah keseluruhan pengamatan dalam dataset Perhitungan nilai akurasi adalah sebagai berikut [25]:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (6)$$

2. Presisi (*precision*)

   Presisi (*precision*) adalah rasio antara pengamatan positif yang diprediksi dengan benar (*True Positive*) dan total pengamatan yang diprediksi sebagai positif (*True Positive* dan *False Positive*). Precision mengukur tingkat akurasi model dalam mengidentifikasi kelas positif. Rumus untuk menghitung nilai presisi adalah sebagai berikut [25]:

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

3. *Recall*

   Recall adalah rasio antara prediksi positif yang benar (*True Positive*) dengan total jumlah pengamatan positif yang sebenarnya ada dalam kelas tersebut (*True Positive* dan *False Negative*). *Recall* mengukur kemampuan model dalam menemukan semua *instance* dari kelas positif yang sesungguhnya. Rumus untuk menghitung nilai *recall* adalah sebagai berikut [25]:

$$Recall = \frac{Tp}{TP+FN} \quad (8)$$

4. *F1-Score*

   F1-Score adalah rata-rata tertimbang antara *precision* dan *recall*. Metrik ini dirancang untuk mempertimbangkan baik *False Positive* (FP) maupun *False Negative* (FN) dalam menilai kinerja model. *F1-Score* memberi bobot lebih besar pada FN dan FP dibandingkan *True Negative* (TN), sehingga TN tidak

memiliki pengaruh signifikan terhadap skor keseluruhan. Rumus untuk menghitung nilai *F1-Score* adalah sebagai berikut:

$$F1 - Score = 2x \frac{recall\ x\ precision}{recall + precision} \quad (9)$$

2.11 *Confusion Matrix*

Evaluasi sistem merupakan komponen penting yang harus dilakukan ketika membuat suatu sistem klasifikasi. Salah satu cara yang dapat dilakukan untuk mengetahui kinerja dan performa model klasifikasi yang telah dibuat adalah dengan menggunakan metode *confusion matrix*. *Confusion matrix* merupakan sebuah tabel yang menunjukkan klasifikasi dari jumlah data uji bernilai benar dan jumlah data uji bernilai salah. *Confusion matrix* terdiri atas matriks dua dimensi, dengan baris dalam matriks tersebut mewakili kelas aktual dari data, sedangkan setiap kolom mewakili kelas prediksi dari data (atau sebaliknya). *Confusion matrix* dapat digambarkan seperti pada tabel dibawah ini [26].

Tabel 1 Tabel *Confusion Matrix*

		Assigned Class	
		Positif	**Negatif**
Actual Class	**Positif**	*True Positive* (TP)	*False Negative* (FN)
	Negatif	*False Positive* (FP)	*True Negative*(TN)

Confusion matrix menggunakan empat istilah untuk merepresentasikan hasil proses klasifikasi, yaitu:

a. *True Positive* (TP), terjadi ketika kelas yang diprediksi positif sesuai dengan fakta yang memang positif.
b. *True Negative* (TN), terjadi ketika kelas yang diprediksi negatif, sesuai dengan fakta yang memang negatif.
c. *False Positive* (FP), terjadi ketika kelas yang diprediksi positif namun faktanya negatif.

d. *False Negative* (FN), terjadi ketika kelas yang diprediksi negatif namun faktanya positif

2.12 Google *Colaboratory*

Google *Colaboratory*, sering disingkat Colab, adalah platform sumber terbuka berbasis *cloud* yang disediakan oleh Google *Research*. Colab merupakan versi replika dari Jupyter Notebook yang menawarkan fasilitas untuk mengolah data dengan teknik *machine learning* atau *deep learning*. *Platform* ini memungkinkan pengguna untuk menulis dan menjalankan kode Python langsung di browser tanpa perlu konfigurasi atau instalasi tambahan. Google Colab memberikan akses mudah ke lingkungan GPU dan memudahkan berbagi file serta kolaborasi dengan pengguna lain, karena terintegrasi langsung dengan Google *Drive* yang beroperasi di sistem *cloud*. Banyak modul yang tersedia di Google Colab, seperti *NumPy*, *SciPy*, *Pandas*, *TensorFlow,* Keras, dan *PyTorch* [27].

2.13 Python

Python adalah bahasa pemrograman tingkat tinggi yang dinamis. Ini adalah bahasa pemrograman yang diinterpretasikan, yang berarti kode sumbernya diubah langsung menjadi kode mesin saat program dijalankan [28]. Python pertama kali diperkenalkan oleh Guido van Rossum pada tahun 1991. Secara umum, Python dapat dijalankan di berbagai sistem operasi dan dapat didistribusikan secara bebas. Dalam fungsinya, Python memanfaatkan berbagai *library* dan *framework* untuk analisis data. *Library* yang ingin diinstal dapat dilakukan menggunakan perintah "pip" [29]. Berikut adalah beberapa *library* yang digunakan dalam penelitian ini:

1. Pandas

Pandas adalah *library* yang menyediakan struktur data dan alat analisis data yang efisien. Untuk memulai, lakukan *import library* dengan menulis import pandas as pd terlebih dahulu. Penggunaan "as" di sini berarti mengganti pemanggilan pandas dengan awalan pd untuk langkah-langkah berikutnya [30].

2. NLTK (*Natural Language Tool Kit*)

Natural Language Toolkit, yang lebih dikenal sebagai NLTK, adalah kumpulan pustaka dan program untuk pemrosesan bahasa alami (NLP) baik secara simbolik maupun statistik, khususnya untuk bahasa Inggris, yang ditulis dalam Python. Tujuan NLTK adalah untuk mendukung penelitian dan pengajaran di bidang NLP atau bidang-bidang terkait seperti linguistik empiris, ilmu kognitif, kecerdasan buatan (*Artificial Intelegence*), pengambilan informasi, dan pembelajaran mesin (*machine learning*). NLTK mendukung berbagai fungsi seperti klasifikasi, tokenisasi, *stemming*, penandaan, penguraian, dan penalaran semantik [31].

3. *Scikit-Learn*

Library *sklearn* adalah *library* python yang digunakan untuk kebutuhan pembelajaran mesin. *Sklearn* mendukung pembelajaran *supervised* dan *unsupervised learning* yang memungkinkan untuk melakukan pekerjaan seperti regresi (*regression*), klasifikasi (*classification*), pengelompokkan/ penggugusan (*clustering*), data *preprocessing*, *dimensionality reduction*, dan model *selection* pembandingan, validasi, dan pemilihan parameter maupun model.

4. *Tweet Harvest*

Tweet-harvest merupakan *tools* yang digunakan untuk melakukan *crawling* data pada media sosial Twitter dengan menggunakan *Application Programming Interface* (API). Dalam penelitian ini *library tweet harvest* digunakan untuk mengambil informasi terkait berita Genosida yang terdapat di X [32].

5. Sastrawi

Sastrawi adalah *library* pada Python yang sederhana dan memberikan peluang kepada pengguna untuk mengonversi kata-kata dalam Bahasa Indonesia dari bentuk infleksi ke bentuk dasarnya (*stem*) [33]. Pada penelitian ini *library* sastrawi digunakan pada proses *pre-processing* data yaitu pada tahapan *stemming* (penghapusan imbuhan).

2.14 Penelitian Terdahulu

Terdapat beberapa penelitian terkait yang dijadikan sebagai perbandingan dan refrensi mengenai metode yang digunakan pada penelitian ini Muhammad Raihan Fais Sya'bani, Uktach Enri, dan Tesa Nur Padilah melakukan penelitian pengklasifikasikan sentimen terhadap beberapa tokoh politik yang berencana mencalonkan diri, yaitu Ganjar Pranowo, Anies Baswedan, Prabowo Subianto, dan Ridwan Kamil, menggunakan algoritma *Naïve Bayes*. Data penelitian diambil dari media sosial Twitter dengan total 3.780 tweet. Hasil akurasi yang didpatkan yaitu Ridwan Kamil 62,5%. Prabowo Subianto 60%, Anies Baswedan 71,43% dan Ganjar 72,68% [34]. Adapun Nanang Husin melakukan penelitian komparasi Algoritma *Naïve Bayes*, BERT dan *Random Forest* untuk *Multi-Class Classification* Pada Artikel *Cable News Network* (CNN). Penelitian ini, metode *naïve bayes classifier* digunakan untuk klasifikasi berbasis pengawasan (*supervised)* dengan memberikan label kelas kepada *instance* menggunakan probabilitas bersyarat, Sedanngkan metode BERT, pada penelitian ini menggunakan BERT *Fine-Tuning*. Untuk mentransfer pengetahuan model yang sudah dilatih untuk dapat menyelesaikan permasalahan lain yang serupa dengan memodifikasi serta memperbaharui parameternya sesuai dengan dataset yang baru. Hasil akurasi yang didapatkan 0.92 dan *macro average f1-score* 0.92 [35].

Pada penelitian yang dilakukan Hendy Syuhada yaitu melakukan pengklasifikasian sentimen terhadap kinerja Komisi Pemberantasan Korupsi pada media sosial twitter. Diperoleh nilai akurasi sebesar 0,64 atau sekitar 64 %. dengan nilai *precision* terbesar didapat pada dataset yang berlabel positif yaitu sebesar 0.69. Nilai *recall* terbesar didapat pada dataset yang berlabel negatif yaitu sebesar 0.89 dari keseluruhan data berlabel negatif. Nilai *F1-score* terbesar didapat pada dataset yang berlabel negatif yaitu sebesar 0.74 [36]. Adapun pada penelitian yang dilakukan Nurfazriah Attamami,Agung Triayudi dan Rima Tamara Aldisa melakukan penelitian performa algoritma klasifikasi *Naïve bayes* dan C4.5, hasil pengujian kinerja kedua model menggunakan *confusion matrix* dengan 730 *record* yang digunakan sebagai data latih dan 313 *record* yang digunakan sebagai data uji,

algoritma klasifikasi C4.5 mendapatkan nilai akurasi yang paling tinggi yaitu sebesar 99.04%. Sebanyak 310 *record* data diprediksi tepat dengan tingkat *error* atau kesalahan sebesar 0.96% atau sebanyak 3 *record* data dari 313 data yang diuji. Sedangkan pada algoritma klasifikasi *Naïve Bayes* di dapatkan nilai akurasi sebesar 92.97%. Sebanyak 291 *record* data diprediksi tepat dengan tingkat *error* atau kesalahan sebesar 7.03% atau sebanyak 22 *record* data diprediksi salah dari 313 data yang dilatih [37]. Kemudian penelitian untuk melakukan analisis sentimen terhadap ulasan film "*Dirty Vote*" pada penelitian ini menunjukkan bahwa model BERT mencapai tingkat kinerja yang tinggi dengan akurasi sebesar 85%, *precision* 86%, *recall* 84% dan *F1-Score* 85% [38].

Selanjutnya terdapat lima penelitian yang membandingkan kinerja *Naïve Bayes* dan BERT dengan berbagai algoritma lain. Penelitian oleh N. Sholihah, dkk., bertujuan untuk menganalisis sentimen publik terhadap kinerja komisi pemilihan umum pasca pemilihan presiden 2024 dengan menggunakan model BERT kemudian diterapkan untuk mengklasifikasikan sentimen komentar dengan kinerja model dievaluasi menggunakan 10-*fold cross validation* hasil evaluasi menunjukkan bahwa lipatan pertama (k-1) mencapai kinerja terbaik dengan akurasi 96%, presisi 96%, *recall* 96% dan *f1-score* 96% [39]. M.Dhito Maulidan, dkk., melakukan penelitian yang menggunakan metode *Naïve bayes* dan metode BERT data yang digunakan sebanyak 3000 data yang dianalisis dengan 1772 *review* positif dan 263 *review* negatif hasil yang didapatkan pada kedua metode ini seimbang yang Dimana menghasilkan akurasi 88,7%, presisi 88,5%, *recall* 100% dan *f1-score* 93,9% [40]. Kemudian D.Sekar,dkk., Melakukan penelitian menggunakan metode *naïve bayes* dengan data pengguna media sosial twitter yang sering menggunakan hastag " *SEA Games* 2023" Hasil dari penelitian ini Sebagian besar tanggapan masyarakat memiliki sentimen positif sebesar 33,4%, sentimen netral 59,1%, sentimen negatif 7,5% kemudian akurasi yang di hasilkan 92,70% [41].

Kemudian penelitian oleh A.Surahman, dkk., melakukan penelitian menggunakan algoritma BERT data yang digunakan berasal dari media sosial Instagram menganail opini publik terhadap produk yang mengalami boikot. Hasil akurasi pada

penelitian ini yaitu McDonald's 84,14%, KFC 95%, Starbucks 94,16%, Burger King 91,42% dan Pizza Hut 93,80% [42]. Penelitian oleh P.A Riyantoko, dkk., melakukan penelitain perbandingan algoritma LSTM dan BERT hasil yang didapatkan yaitu pada LSTM akurasi 98,22%, *precision* 94,41%, *recall* 92,98% dan *F1-Score* 92,62%. Kemudian hasil yang didapat pada BERT, akurasi 99,35%, *precision* 97,87%, *recall* 96,51% dan *F1-Score* 94,55% [43]. Kemudian penelitian A.Fauzi menggunakan *Random Forest* untuk Sentimen Bahasa Indonesia dengan *GridSearch dan SMOTE* dengan data sebanyak 611 dari media sosial X hasil yang didapatkan sebesar 89% [44].

Tabel 2 Penelitian Terdahulu

No	Peneliti	Algoritma	Dataset	Hasil
1.	M.R.F Sya'bani, dkk (2022) [34].	*Naïve Bayes*	Dataset yang diperoleh melalui twitter dengan menggunakan hastag: #capres2024 #ganjarpranowo #aniesbaswedan #prabowo #ridwankamil	Akurasi Ridwan.K:62,5%. Prabowo:60%, Anies.B:71,43% Ganjar: 72,68%
2.	N.Husin(2023) [35].	-*Naïve Bayes* -BERT -*Random Forest*	Data artikel berita *Cable News Network* (CNN) dari tahun 2011 sampai 2022	BERT Akurasi : 92% *F1-Score* : 92%
3.	N.Atamimi, dkk (2023) [37].	-*Naïve Bayes* -C4.5	Dataset yang diperoleh dari Dinas Kesehatan Kota Depok, tahun 2020-2022.	Naïve Bayes Akurasi : 9297% Presisi : 61,80% *Recall* : 89,55% *F1-Score* : 73,17 C4.5 Akurasi: 99,04% Presisi : 57,74% *Recall* : 99,44% *F1-Score*: 73,06%
4.	D.Sjoraida, dkk (2024) [38].	BERT	Dataset yang diperoleh yaitu dari Kumpulan data mengenai ulasan "*Dirty Vote*"	Akurasi : 85% Presisi : 86% *Recall* : 84% *F1-Score* : 85%

No	Peneliti	Algoritma	Dataset	Hasil
5.	N.Sholihah, dkk (2024) [39].	BERT	Dataset Komen KPK di Media Youtube.	10 pengujian dengan nilai K-1 menghasilkan: Akurasi: 96% Presisi :96% *Recall* : 96% *F1- Score* :96%
6.	M. Maulidan, dkk (2024) [40].	-*Naïve Bayes* -BERT	Dataset yang diperoleh dari *Review* Aplikasi Simple pol pada Google *Playstore*	Hasil yang dihasilkan sama imbang antara *Naïve bayes* dan BERT yaitu: Akurasi: 88,7% Presisis:88,5% *Recall*:100% *F1-Score*:93,9%
7.	D.Arum, dkk (2023) [41].	*Naïve Bayes*	Dataset yang diperoleh tanggapan mengenai Sea Game 2023 pada media sosial yaitu Twitter(X).	Dilakukan 5 kali pengujian dan hasil akurasi dengan rasio 40:60 menghasilkan akurasi 92,70%
8.	A.Sulaeman, dkk (2024) [42].	BERT	Dataset yang diperoleh yaitu komentar terkait produk-produk boikot seperti: McDonalds, Sturbucks KFC, Burger King, Pizza Hut pada media sosial instagram.	Hasil akurasi pada produk boikot yaitu: KFC:95% McDonald: 84,14% Starbucks: 94,16% BurgerKing: 91,42% PizzaHut: 93,80%

No	Peneliti	Algoritma	Dataset	Hasil
9.	P.A.Riyantoko, dkk (2022) [43].	-BERT -LSTM	Data dari UCI – *Machine Learning* " SMS *Spam Collection* data set" sebanyak 5572 SMS	BERT Akurasi :99,35% Presisi:97,87% *Recall*:96,51% F1-Score:94,55% LSTM Akurasi:98,22% Presisi:94,41% *Recall*: 92,98% *F1-Score*:92,62%
10.	H.Syuhada (2022) [36].	*Naïve Bayes*	Data yang diperoleh melalui Twitter hastag #KPK	Akurasi:64% *Precision*:0.69 *Recall*:0.89. *F1-score* :0.74
11.	A.Fauzi, dkk (2025) [44].	*Random Forest + Hyperparameter Gridsearch*	Data yang diperoleh melalui media sosial X dengan #UKT data didapat sebanyak 611 data.	Hasil akurasi *cross validation* optimasi algoritma *random forest* dengan *hyperparameter gridsearch* 89%

Pada penelitian ini akan dilakukan pengklasifikasian terhadap *post* berita Genosida di media sosial X dengan menggunakan Algoritma BERT dan *Naïve Bayes Classifier*. BERT dikenal dengan efisiensi pelatihannya yang tinggi serta kemampuannya menghasilkan performa unggul dalam berbagai penelitian terkait klasifikasi. Di sisi lain, *Naïve Bayes Classifier* populer karena kemudahan dan kecepatannya dalam memprediksi kelas pada kumpulan data uji. Kemudian Evaluasi pada penelitian ini dilakukan menggunakan *classification report* (*recall, precision, f1-score, accuracy*), dan *confusion matrix*.

III. METODE PENELITIAN

3.1 Waktu dan Tempat Penelitian

Waktu dan tepat penelitian dilakukan pada:

1. Waktu Penelitian : Desember 2024 sampai dengan September 2025
2. Tempat Penelitian : Laboraturium Terpadu Jurusan Teknik Elektro Universitas Lampung

Tabel 3.Waktu Penelitian

No	**Kegiatan**	**Waktu Pelaksanaan Penelitian**									
		2024	2025								
		des	Jan	feb	mar	Apr	Mei	Jun	jul	agt	sep
1	Studi Literasi										
2	Persiapan Alat dan Bahan										
3	Business *Understanding*										
4	Data *Understanding*										
5	Data *Preparation*										
6	*Modelling*										
7	*Evaluation*										
8.	Analisis Hasil										

3.2 Alat dan Bahan Penelitian

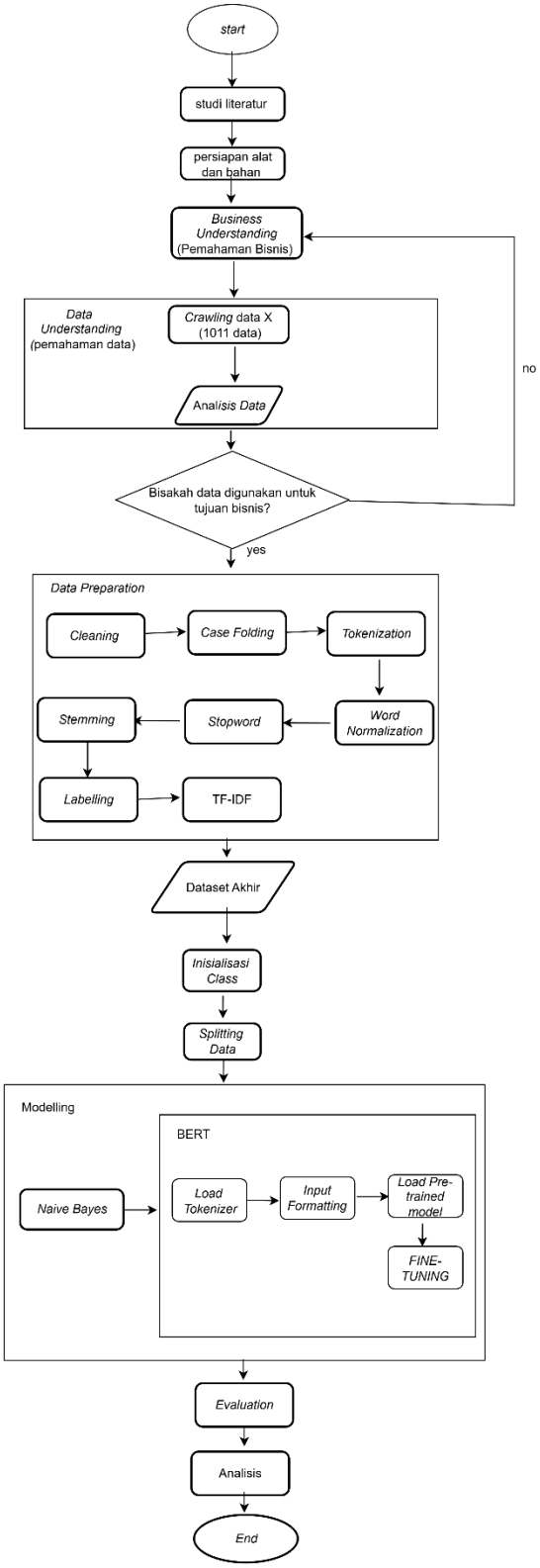
Adapun alat dan bahan dari penelitian ini sebagai berikut:

Tabel 4. Alat dan Bahan Penelitian

No	Nama	Spesifikasi	Kegunaan
1.	Laptop	Vivobook_Asus Laptop AMD *Ryzen* 3 3250U *with Radeon Graphics*	Perangkat untuk pengembangan aplikasi
2	Python	Google *Colaboratory on* Python 3.11.2	Sebagai bahasa pemrograman untuk membangun model.
3.	VS Code	Versi 1.76.0	*Text Editor*
4.	Windows	Windows 11	Sebagai sistem operasi

3.3 Tahapan Penelitian

Metodologi yang digunakan dalam penelitian ini adalah metode *Cross Industry Standard Process for Data Mining* (CRISP-DM) dan algoritma pengklasifikasian teks seperti BERT dan *Naïve Bayes Classifier*. CRISP-DM menawarkan proses yang terstruktur untuk menyelesaikan masalah dalam unit bisnis atau penelitian dengan menggunakan teknik data mining yang tepat. Metode CRISP-DM ini terdiri dari enam tahapan, namun pada penelitian ini hanya dilakukan sampai lima tahap seperti yang ditunjukkan pada diagram alur di bawah ini.



Gambar 3.1 *Flowchart* Tahapan Penelitian

Berdasarkan dari diagram alur dari metode CRISP-DM tahapan yang dilakukan dalam penelitian ini, yaitu sebagai berikut:

3.3.1 *Business Understanding* (Pemahaman Bisnis)

Pada tahap ini dilakukan beberapa langkah, seperti mengidentifikasi tujuan bisnis, menilai kondisi, serta menetapkan tujuan untuk data mining. Tujuan penelitian ini adalah membangun sistem analisis sentimen berdasarkan unggahan masyarakat di *platform* X yang membahas isu genosida, yang diklasifikasikan ke dalam sentimen positif, netral, dan negatif. Genosida yang dimaksud dalam penelitian ini merujuk pada upaya penghentian genosida, khususnya pada konflik kemanusiaan yang tengah terjadi di Palestina. Penelitian ini bertujuan melihat bagaimana opini masyarakat terhadap isu tersebut dibagikan di media sosial. Dalam konteks ini, dilakukan pengkategorian sentimen, yaitu, sentimen negatif, apabila seseorang mengungkapkan pendapat dengan kata-kata yang tidak baik atau bernada buruk. Kemudian Sentimen positif, apabila unggahan berisi dukungan dengan kata-kata yang baik dan pantas. Kemudian sentimen netral, mengarah pada informasi yang bersifat objektif, tidak menyudutkan pihak tertentu, dan tidak mengandung ujaran kebencian. Pada penelitian ini, pengklasifikasian sentimen awalnya dilakukan secara manual, yang memerlukan waktu cukup banyak dan tidak efisien. Oleh karena itu, tujuan data mining dalam tahap ini adalah melakukan klasifikasi sentimen secara otomatis menggunakan algoritma BERT dan *Naïve Bayes Classifier,* kemudian menganalisis kinerja dari kedua algoritma tersebut,

3.3.2 Data *Understanding* (Pemahaman Data)

Pada tahapan dilakukan pemahaman terhadap kebutuhan data sesuai dengan tujuan bisnis pada tahap sebelumnya dan mengumpulkan dataset yang digunakan untuk tahap selanjutnya. Dimana dalam penelitian kali ini, dataset yang digunakan adalah *Posting* dari masyarakat yang diperoleh melalui media sosial X. Data *posting* tersebut diambil melalui proses *crawling* data dengan *library tweet harvest* yang

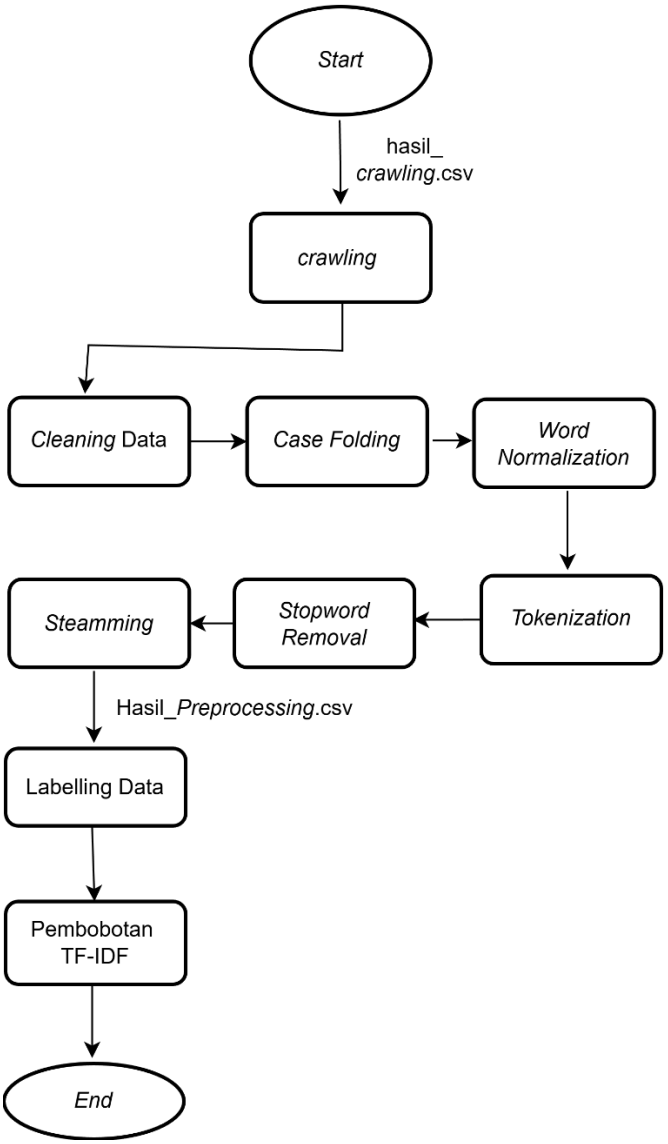
ada pada python. Dimana data yang diambil melalui X ini merupakan data pada bulan Oktober 2023 sampai dengan Desember 2024 dengan jumlah data sebanyak 1011 data. Dataset yang didapatkan pada tahap ini disimpan dalam bentuk csv. Selanjutnya dataset tersebut diolah dan dianalisa agar didapatkan informasi**.**

created_at	username	full_text
Mon Dec 30	VirGo04897515	@SiregarMaralel4 @vianratulangi Lu yg genosida bangsat hamas! Tolol!!
Mon Dec 30	a_daud_ppnp	Pakar: Dunia Hanya Bisa Menyaksikan Genosida di Gaza dengan Ketidakberdayaan Total! https://t.co/mlmfo5gQng
Mon Dec 30	Basheera_Putry	Aktivis mengorganisir demonstrasi di Swedia membawa boneka yang meniru tubuh anak-anak martir yang dibunuh oleh pendudukan dalam perang
Mon Dec 30	yosunoku	@tanyarlfes Pemuja isnotreal = pendukung genosida
Mon Dec 30	kaminarinafas	@ardi_tama1 Genosida secara ga langsung wkwkwk
Mon Dec 30	DjajaSampurna	@KompasTV Harta pak presiden yang katanya trilyunan rupiah aoakah didapat dengan cara legal atau illegal? Kalau pejabat negara termssuk presi
Mon Dec 30	DjajaSampurna	@BosPurwa @bank_indonesia @DPR_RI Bagi bagi jatah maling. Negeri ini butuh program bunuh koruptor. Segerakan revolusi dan genosida korupt
Mon Dec 30	furisenna	Genosida tp halus..
Mon Dec 30	chatnoir7777	Genosida terjadi setiap hari di Gaza. Sarah ingin bertahan hidup bersama keluarganya. Sarah dan keluarganya berada di ambang kematian. Semua
Mon Dec 30	heruditabuti	@arasharapova @dikaizm Honestly that s fucked up. Gak ada satupun manusia di dunia ini yang patut hidup dalam genosida termasuk di masa naz
Mon Dec 30	amareguzel	@angels_bac58190 itu bagian filter nya x p4j33t (india jikonis) mereka ada dibelakang akun2 pro palestine yg disuspen akun aku yg @uniqueofme1
Mon Dec 30	ag_tr67712	@Willded1 @UstadzWaloni Daftar pembantaian dan genosida domba2 tersesat https://t.co/rgleHeUFze
Mon Dec 30	TRSP317	@Willded1 @UstadzWaloni Manusia kemakan propaganda mamarika buka mata lebih jauh genosida isrewel ke gaza itu siapa Agamanya dan siapa
Mon Dec 30	SiregarMaralel4	@vianratulangi Terroris itu seperti kau jelas kejadian di Palestina itu Genosida masih kau bilang terroris RI jelas konstitusi melarang atribut yg berba
Mon Dec 30	pochettenooutaa	@haitanizu Iyaa dia tuh sebetulnya teguh pendirian. Sayangnya sama aja dengan menyetujui genosida â•¥ĩâ•¥
Mon Dec 30	NYIRATUKIDULL	Genosida yang disebarkan Netanyahu yang mendapat tepuk tangan dari AS dan Eropa menjadi sebuah undangan bagi barbarisme global. Sekarang
Mon Dec 30	ismailabir94	@DaveMilbo Kepala bapa anda tiada genosida @MastSydney
Mon Dec 30	Taqwasquad	Seorang aktivis London memasuki McDonald's dan mengingatkan tentang genosida di Gaza https://t.co/ozHy5HIieQ
Mon Dec 30	Pribumildn	Ini bukan Gaza tapi Kota Der EZ- Zor Suriah akibat genosida oleh rezim Assad Rusia Hizbullat dan Iran. https://t.co/5BgP6kvbsH
Mon Dec 30	NYIRATUKIDULL	Bagikan klipnya ke mana saja Memukuli dan membunuh warga Suriah yang tidak berdaya Di Suriah pada 29/12/2024 Genosida terhadap warga Suri
Mon Dec 30	PHeruchakra	@GoodNewsJurnal Ajaran setan itu yg mendukung LGBT... Smua dihalalkan... Hobi perang... Ngebajak bangsa lain.... Genosida...
Mon Dec 30	erlanishere	Di Gaza sedang terjadi cuaca ekstrem dan krisis lingkungan akibat genosida berkepanjangan. Warga yang tinggal di tenda-tenda pengungsian sering
Mon Dec 30	NiaArmiani17	@BlacPitt2 @GoodNewsJurnal @erlanishere Yg setan itu Setanyahu dan para zionajis pelaku genosida yg jdi buronan internasional itu. Kalo lu kete
Mon Dec 30	eka_gendut	@DeningCarlo Tak beda dengan pencuri berseragam.. udahlah. Dunia ini sudah jatuh parah. Terima saja genosida yg akan datang ini...
Mon Dec 30	Putralaut_	@dina_sulaeman Teriris campur geram rasanya hati setiap mlht barak pengungsi aplgi RS yg di hancurkan sampai kpn ini hrs dibiarkan pdhl sdh jlsÂ

Gambar 3.2 Dataset yang didapatkan pada saat *Crawling* data di X.

3.3.3 Data *Preparation* (Persiapan Data)

Pada tahapan persiapan data proses yang dilakukan adalah mengolah dataset akhir yang sebelumnya telah diperoleh untuk nantinya dimasukkan ke tahap selanjutnya yaitu tahap pemodelan. Untuk mengolah dataset tersebut dilakukan proses *preprocessing* dan pembobotan TF-IDF. Tahapan *pre-processing* mencakup berbagai proses seperti *cleaning*, *case folding*, *tokenization*, *filtering* (*stopword removal*). Selanjutnya setelah dilakukan proses *pre-processing* pada dataset maka dilakukan proses *labelling* data yang kemudian disimpan kedalam dataset untuk dimasukkan ke tahap pemodelan. Pada penelitian kali ini akan menggunakan 3 label sentimen yaitu positif, netral, dan negatif. Pada pelebelan data Kemudian setelah data telah diberi label maka tahap selanjutnya yaitu melakukan pembobotan kata dengan menggunakan metode pembobotan kata TF-IDF. Berikut Daiagram alur dari tahapan persiapan data ditunjukkan pada gambar 3.3.



Gambar 3.3 Diagram Alur Persiapan Data

a. *Cleaning* Data

Pada tahap ini membersihkan data dari karakter-karakter yang tidak diperlukan seperti tautan, nama pengguna, angka, simbol, dan kata-kata yang mengganggu proses analisis. Berikut merupakan salah satu contoh dari cleaning data dengan menggunakan data *crawl* yang dapat dilihat pada tabel 5.

Tabel 5. Contoh dari *Cleaning* data

No	Input	Output
1.	Seorang teknisi Google melakukan protes di sebuah konferensi teknologi di New York. Dia menuduh seorang eksekutif dari operasi raksasa pencarian yang berbasis di Israel mendukung genosida dengan bekerja sama dengan pemerintah Israel. Barak Regev direktur pelaksana. Google https://t.co/UZPyoLnSmE	Seorang teknisi Google melakukan protes di sebuah konferensi teknologi di New York Dia menuduh seorang eksekutif dari operasi raksasa pencarian yang berbasis di Israel mendukung genosida dengan bekerja sama dengan pemerintah Israel Barak Regev direktur pelaksana Google

b. *Case Folding*

Pada tahap ini dilakukan untuk mengubah teks menjadi huruf kecil atau huruf besar sesuai kebutuhan klasifikasi, Berikut merupakan salah satu contoh dari *Case Folding* dengan menggunakan data *crawl* yang dapat dilihat pada tabel 6.

Tabel 6 Contoh *Case Folding*

No	Input	Output
1.	Seorang teknisi Google melakukan protes di sebuah konferensi teknologi di New York. Dia menuduh seorang eksekutif dari operasi raksasa pencarian yang berbasis di Israel 'mendukung genosida' dengan bekerja sama dengan pemerintah Israel. Barak Regev direktur pelaksana Google https://t.co/UZPyoLnSmE	seorang teknisi google melakukan protes di sebuah konferensi teknologi di new york dia menuduh seorang eksekutif dari operasi raksasa pencarian yang berbasis di israel mendukung genosida dengan bekerja sama dengan pemerintah israel barak regev direktur pelaksana google

c. *Tokenization*

Pada *Tokenization* dilakukan memecah deskripsi yang awalnya berupa kalimat menjadi kata-kata. Berikut merupakan contoh dari *Tokenization* dengan menggunakan data *crawl* yang dapat di lihat pada tabel 7.

Tabel 7. Contoh dari *Tokenization*

No	Input	Output
1.	seorang teknisi google melakukan protes di sebuah konferensi teknologi di new york dia menuduh seorang eksekutif dari operasi raksasa pencarian yang berbasis di israel mendukung genosida dengan bekerja sama dengan pemerintah israel barak regev direktur pelaksana google.	[‘seseorang’,‘teknisi’,‘google’, ‘melakukan’, ‘protes’, ‘di’, ‘sebuah’, ‘konferensi’, ‘teknologi’, ‘di’, ‘new’, ‘york’, ‘dia’, ‘menunduh’, ‘seorang’, ‘eksekutif’, ‘dari’, ‘operasi’, ‘raksasa’, ‘pencarian’, ‘yang’, ‘berbasis’, ‘di’, ‘israel’, ‘mendukung’, ‘genosida’, ‘dengan’, ‘bekerja’, ‘sama’, ‘dengan’, ‘pemerintah’, ‘israel’, ‘barak’, ‘regev’, ‘direktur’, ‘pelaksana’, ‘google’.]

d. *Word Normalization*

Pada tahap ini dilakukan mengoreksi kata-kata yang tidak tepat dengan menggunakan kamus agar memiliki arti yang sesuai, sehingga tidak menambah dimensi vektor yang bisa memperlambat proses komputasi. Berikut merupakan contoh dari *word normalization* dengan menggunakan data *crawl* yang dapat dilihat pada tabel 8.

Tabel 8 Contoh dari *Word Normalization*

No	Input	Output
1.	Seorang teknisi Google melakukan protes di sebuah konferensi teknologi di New York. Dia menuduh seorang eksekutif dari operasi raksasa pencarian yang berbasis di Israel 'mendukung genosida' dengan bekerja sama dengan pemerintah Israel. Barak Regev direktur pelaksana Google https://t.co/UZPyoLnSmE	seorang teknisi google melakukan protes di sebuah konferensi teknologi di new york dia menuduh seorang eksekutif dari operasi raksasa pencarian yang berbasis di israel **mendukung genosida** dengan bekerja sama dengan pemerintah israel barak regev direktur pelaksana google.

e. *Stopword Removal*

Pada tahap ini dilakukan menghapus kata-kata yang sering muncul tetapi tidak relevan, seperti kata hubung yang tidak berpengaruh pada proses klasifikasi, berikut

merupakan contoh dari *stopword removal* dengan menggunakan data *crawl* yang dapat dilihat ditabel 9.

Tabel 9 Contoh *Stopword Removal*

No	Input	Output
1.	"Seorang" teknisi Google "melakukan" protes "di sebuah" konferensi teknologi "di" New York. "Dia" menuduh "seorang" eksekutif "dari" operasi raksasa pencarian "yang" berbasis di Israel mendukung genosida "dengan" "bekerja sama dengan" pemerintah Israel. Barak Regev direktur pelaksana Google	['teknisi', 'google', 'protes', 'konferensi', 'teknologi', 'new', 'york', 'menuduh', 'eksekutif', 'operasi', 'raksasa', 'pencarian', 'berbasis', 'israel', 'mendukung', 'genosida', 'pemerintah', 'israel', 'barak', 'regev', 'direktur', 'pelaksana', 'google'.]

f. *Stemming*

mengubah kata-kata dalam bahasa Indonesia menjadi bentuk dasarnya dengan menghilangkan imbuhan awalan, akhiran, dan sisipan. Berikut merupakan scontoh pada *stemming* dengan menggunakan data *crawl* yang dapat dilihat pada tabel 10.

Tabel 10 Contoh *Stemming*

No	Input	Output
1.	Teknisi google protes konferensi teknologi new York **me**nuduh eksekutif operasi raksasa **pen**cari**an** **ber**basis israel **men**dukung genosida pemerintah israel barak regev direktur **pe**laksana google	teknisi google protes konferensi teknologi new york **tuduh** eksekutif operasi raksasa **cari** **basis** israel **dukung** genosida perintah israel barak regev direktur **laksana** google.

g. *Labelling* Data (Pelabelan Data)

Labelling data tersebut dilakukan untuk memberikan label atau kelas pada data *post* hasil *crawling* data yang sebelumnya telah diperoleh. Selanjutnya hasil *labelling* tersebut dimasukkan pada dataset dan selanjutnya digunakan pada proses pemodelan. Terdapat tiga label yang digunakan untuk analisis sentimen ini yaitu

positif, netral, dan negatif. Proses pelabelan ini dilakukan secara manual dengan pengkategorian kata untuk sentimen dilakukan berdasarkan konotasi atau makna umum *terkait* dengan kata-kata tersebut. Berikut merupakan pengkatagorian dari *Labelling*.

Tabel 11 Pengkategorian Label

No	Sentimen	Pengkategorian
1.	Sentimen Negatif	1. Apabila orang menghujat dengan kalimat negatif seperti: berkata kasar dan mengeluarkan perkataan yang menggunakan Bahasa kurang baik contohnya *post* yang terdapat kemarahan dengan menggunakan kata Binatang dan tidak seharusnya. Contoh: "Genosida bangsat hamas tolol" 2. Apabila orang mempropokatif untuk melakukan serangan genosida. Contohnya: "Apalagi pejabat hukuman nya minimal mati atau keluarganya di genosida habis" 3. Apabila orang menyatakan dan menggiring opini yang tidak jelas kebenarannya. Contohnya:" Umat Islam sedang dibantai tapi dunia diam saja. Karena pembantainya adalah sekutu mereka"
2.	Sentimen Positif	1. Apabila orang berkata dengan unsur mendukung, Tindakan solutif atau pemulihan, serta apresiasi dan penghargaan dalam menangani genosida Contohnya: "Di Oscar sekitar 400 selebriti tergabung dalam gerakan Artists4Ceasefire untuk menyerukan gencatan senjata di konflik" 2. Apabila orang yang berpihak terhadap korban genosida, seperti memberi semangat dan dukungan serta melakukan aksi pemboikitan. Contohnya: "Siap boikot product product Turki di Indo kalau mereka tetap mendukung genosida"

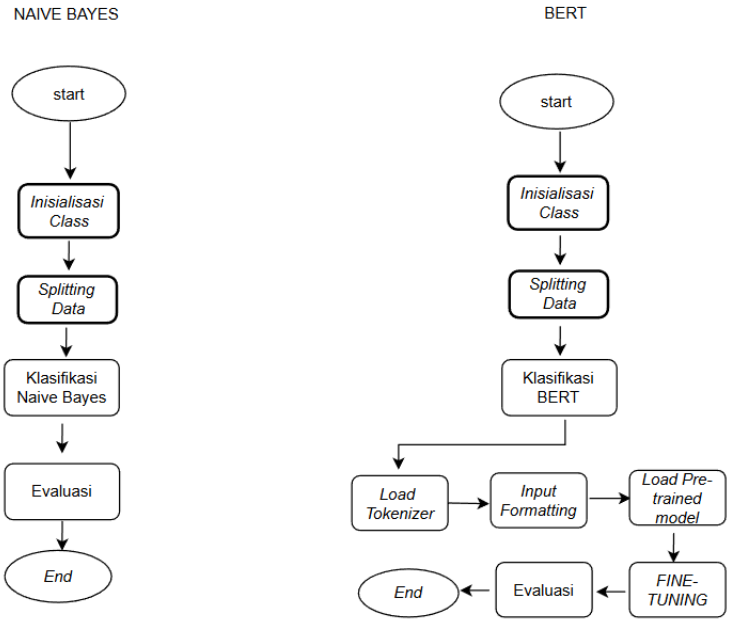
3.	Sentimen Netral	Apabila orang berkata dengan unsur kata-kata yang tidak secara khusus mengekspresikan perasaan atau emosi serta tidak berpihak pada sesuatu, melainkan lebih mengarah kepada deskripsi umum atau informasi factual. Contoh: "Pakar Dunia Hanya Bisa Menyaksikan Genosida di Gaza dengan Ketidakberdayaan Total! https://t.co/mlmfo5gQng"

3.3.4 *Modelling* (Pemodelan)

Pada tahap pemodelan, penelitian ini menggunakan dua algoritma, yaitu BERT dan *Naïve Bayes classifier*. Setelah seluruh data terkumpul, dataset terlebih dahulu dibagi menjadi dua bagian sebelum masuk ke proses pelatihan model, yaitu data *training* (latih) dan data *testing* (uji). Pembagian dilakukan dengan rasio 80:20, di mana 80% data digunakan sebagai data latih dan 20% sisanya sebagai data uji. Proses pelatihan model memanfaatkan data latih tersebut agar model mampu mengenali pola sebelum dievaluasi menggunakan data uji. Dalam penelitian ini, algoritma *Naïve Bayes* yang digunakan adalah *Multinomial Naïve Bayes*, karena jenis ini sangat sesuai untuk data berbasis teks. *Multinomial Naïve Bayes* bekerja dengan menghitung frekuensi kemunculan kata dalam dokumen, sehingga algoritma ini efektif untuk permasalahan klasifikasi teks seperti analisis sentimen. Model ini mengasumsikan bahwa fitur berupa kata bersifat *multinomial* (berdasarkan jumlah/frekuensi), yang membuatnya lebih tepat digunakan dibanding *Gaussian* atau *Bernoulli Naïve Bayes* untuk data teks. Sementara itu, pada pemodelan menggunakan BERT, proses pelatihan dilakukan dengan menerapkan *batch size* dan *epoch*. Penelitian ini menggunakan *batch size* sebesar 32 dan jumlah *epoch* sebanyak 3. *Batch size* 32 dipilih karena memberikan keseimbangan antara kebutuhan memori dan stabilitas pembelajaran; *batch* yang lebih kecil membantu mengurangi beban GPU dan membuat proses pelatihan lebih efisien.

Adapun cara kerja *batch size*, yaitu jumlah sampel yang diproses sekaligus sebelum model memperbarui parameter. Dengan *batch size* 32, berarti model memproses 32

kalimat setiap kali melakukan satu langkah pembaruan. *Batch* kecil seperti ini sering menghasilkan gradien yang cukup stabil tanpa membutuhkan memori yang besar. Sedangkan *epoch* menggambarkan berapa kali seluruh dataset pelatihan diproses oleh model secara penuh. Dengan menggunakan 3 *epoch*, artinya seluruh data latih akan dilalui sebanyak tiga kali selama proses *training*. Pemilihan 3 *epoch* dilakukan agar model memperoleh pemahaman yang cukup terhadap pola dalam data tanpa mengalami *overfitting* atau pelatihan berlebihan. Secara keseluruhan, berikut ini merupakan tahapan alur kerja dari algoritma BERT dan *Multinomial Naïve Bayes* pada penelitian ini.



Gambar 3.4 *Flowchart Naïve Bayes* dan BERT

Gambar 3.4 memperlihatkan perbandingan alur proses klasifikasi sentimen menggunakan algoritma *Naïve Bayes* dan BERT. Pada metode *Naïve Bayes*, proses dimulai dari inisialisasi *class*, diikuti oleh *splitting* data untuk membagi data latih dan uji. Selanjutnya dilakukan klasifikasi *Naïve Bayes*, dan hasilnya dievaluasi. Sementara itu, alur BERT memiliki tahapan yang lebih kompleks. Setelah inisialisasi *class* dan *splitting* data, dilakukan proses klasifikasi BERT. Untuk memproses data teks, digunakan *tokenizer* yang mengubah teks menjadi format numerik (*input formatting*). Data ini kemudian dimasukkan ke dalam *pre-trained*

model BERT, yang selanjutnya dapat di *fine-tune* sesuai data latih yang digunakan. Proses ini diakhiri dengan tahap evaluasi untuk mengukur performa model.

3.3.5 *Evaluation* (Evaluasi)

Setelah tahap pengklasifikasi sentimen dengan menggunakan Algoritma BERT dan *naïve bayes classifier* telah selesai dilakukan, tahapan berikutnya adalah melakukans evaluasi terkait model yang telah dibuat pensgklasifikasian. Evaluasi pada penelitian ini dilakukan menggunakan *classification report (recall, precision, f1-score, accuracy),* dan *confusion matrix.*

V. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Adapun kesimpulan yang diperoleh berdasarkan hasil penelitian yang telah dilakukan adalah sebagai berikut:

1. Berdasarkan hasil evaluasi terhadap 1200 data, algoritma *Naïve Bayes Classifier* menunjukkan akurasi lebih tinggi yaitu 81% dibandingkan algoritma BERT yang memperoleh akurasi 73%. Namun, jika dilihat dari metrik *precision*, *recall*, dan *F1-Score*, BERT memberikan performa yang lebih baik dan lebih konsisten antar kelas, dengan rata-rata perbandingan *precision* 0,80, *recall* 0,73, dan *F1-Score* 0,74. Pada kelas negatif, *Naïve Bayes* unggul dengan *precision* 1,00, tetapi pada kelas netral BERT jauh lebih unggul dengan *precision* 0,93 dan *F1-Score* 0,76, menunjukkan kemampuannya dalam memahami konteks yang ambigu. Dengan demikian, meskipun akurasi total BERT lebih rendah, performanya lebih stabil dan seimbang, sehingga lebih mampu menangani variasi bahasa dan kompleksitas konteks dibandingkan *Naïve Bayes.*

2. Pada evaluasi kinerja model, menggunakan metode evaluasi berbasis *confusion matrix* dengan empat metrik utama, yaitu akurasi, *precision*, *recall*, dan *F1-score*. Proses implementasi menggunakan Python dengan bantuan *library* seperti *scikit-learn* dan *Hugging Face Transformers*. Berdasarkan hasil evaluasi yang didapatkan model BERT direkomendasikan untuk analisis sentimen ulasan Genosida karena mampu menangkap konteks kalimat yang lebih kompleks serta menunjukkan performa yang lebih konsisten antar kelas. Meskipun demikian, baik *Naïve Bayes* maupun BERT sama-sama memiliki

kemampuan yang baik dalam melakukan pengklasifikasian sentimen, dan perbedaan akurasi keduanya tidak terlalu jauh sehingga keduanya tetap layak digunakan sesuai kebutuhan penelitian.

5.2 Saran

Adapun saran yang dapat diberikan untuk penelitian selanjutnya berdasarkan penelitian yang telah dilakukan adalah sebagai berikut:

1. Jumlah pelabelan sentimen pada dataset sebaiknya memiliki jumlah yang seimbang atau proposional sehingga akan diperoleh nilai akurasi model yang lebih optimal.

2. Pemodelan klasifikasi terhadap dataset Genosida dapat dilakukan menggunakan algoritma lain seperti *Random Forest* yang dioptimalkan dengan teknik *GridSearch* dan *SMOTE*, agar dapat diketahui metode mana yang mampu memberikan performa klasifikasi yang lebih baik pada dataset dengan jumlah data yang terbatas.

DAFTAR PUSTAKA

[1] M. H. Prasetyo, "Kejahatan Genosida Dalam Perspektif Hukum Pidana Internasional," *Gema Keadilan*, vol. 7, no. 3, pp. 115–138, 2020, doi: 10.14710/gk.2020.9075.

[2] S. G. Vanya Karunia Mulia Putri, "Pengertian Kejahatan Genosida dan Contohnya," kompas.com. [*Online*]. *Available*: https ://www.kompas.com/ /read/2021/07/26/141814669/pengertian-kejahatan-genosida-dan contohnya

[3] H. Husna, "Pemanfaatan *Platform* X Dalam Penyebaran Berita Oleh Media GoRiau.com," Riau, 2025. [*Online*]. *Available*: https://repository.uin-suska .ac.id/90280/1/Skripsi Gabungan - Hanifatul Husna Ilmu Komunikasi.pdf

[4] F. Hasan, "Kenapa Twitter Jadi X? Ini Penjelasan dan Perubahan Fiturnya, " detik.net. *Accessed*: Nov. 17, 2024. [*Online*]. Available: https://i net.detik .com/cyberlife/d-6869515/kenapa-twitter-jadi-x-ini-penjelasan-dan-perubahan-fiturnya

[5] S. N. Dhage and C. K. Raina, "A review on Machine Learning Techniques," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 4, no. 3, pp. 395–399, 2016, [Online]. Available: http://www.ijritcc.org

[6] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Comput. Intell. Mag.*, vol. 9, no. 2, pp. 48–57, 2014, doi: 10.1109/MCI.2014.2307227.

[7] R. Feldman and J. Sanger, *The text mining handbook : advanced approaches in analyzing unstructured data*. Cambridge, England: Cambridge University Press, 2007.

[8] W. Khan, A. Daud, K. Khan, S. Muhammad, and R. Haq, “Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends,” *Nat. Lang. Process. J.*, vol. 4, p. 100026, Sep. 2023, doi: 10.1016/j.nlp.2023.100026.

[9] A. Sasmita, G. A. Pradnyana, and D. G. H. Divayana, “Pengembangan Sistem Analisis Sentimen Untuk Evaluasi Kinerja Dosen Universitas Pendidikan Ganesha Dengan Metode Naive Bayes,” *JST (Jurnal Sains dan Teknol.*, vol. 11, no. 2, pp. 451–462, Sep. 2022, doi: 10.23887/ jstundiksha. v11i2.44384.

[10] U. G. Jing Ge, Marisol Alonso Vazquez, “Sentiment analysis: a review. In Sigala, M. & Gretzel, U. (Eds.), Advances in Social Media for Travel, Tourism and Hospitality,” *Adv. Soc. Media Travel*, pp. 243–261, 2018.

[11] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Minneapolis, Minnesota, USA: Association for Computational Linguistics (ACL), Oct. 2019, pp. 4171–4186. [Online]. Available: http://arxiv.org/abs/1810.04805

[12] Susan, “Gambar Arsitektur BERT,” sushant-kumar. Accessed: Nov. 25, 2025. [Online]. Available: https://sushant-kumar.com/blog/bert

[13] F. A. Furfari, “The Transformer,” *IEEE Ind. Appl. Mag.*, vol. 8, no. 1, pp. 8–15, 2002, doi: 10.1109/2943.974352.

[14] Trivusi, “Apa Bedanya *Epoch* dan *Batch Size* pada Deep Learning?,” trivusi.web. Accessed: Nov. 01, 2025. [*Online*]. *Available*: https:/ /www. trivusi.web.id/2022/08/epoch-dan-batch-size.html?m=1

[15] S.Ananta, “How to choose Batch Size and Number of Epochs When Fitting a Model?,” geeksforgeeks. Accessed: Nov. 01, 2025. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/how-to-choose-batch-size-and-number-of-epochs-when-fitting-a-model/

[16] A. R. T. Lestari, R. S. Perdana, and M. A. Fauzi, "Analisis Sentimen Tentang Opini Pilkada Dki 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naïve Bayes dan Pembobotan Emoji," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. December, pp. 1718–1724, 2017, [Online]. Available: http://j-ptiik.ub.ac.id

[17] F. Handayani, D. Feddy, and S. Pribadi, "Implementasi Algoritma Naive Bayes Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110," *J. Tek. Elektro*, vol. 1, no. 1, pp. 19–24, 2015.

[18] S. Natalius, "Metoda Naïve Bayes Classifier dan Penggunaannya pada Klasifikasi Dokumen," Bandung, 2020. [Online]. Available: https://adoc. pub/metoda-nave-bayes-classifier-dan-penggunaannya-pada-klasifik.html# google_vignette

[19] P. Chapman *et al.*, "Crisp-Dm," *SPSS inc*, vol. 78, pp. 1–78, 2000, [Online]. Available: http://www.crisp-dm.org/CRISPWP-0800.pdf

[20] Y. Findawati and M. A. Rosid, *Buku Ajar Text Mining*. Sidoarjo, Jawa Timur: Umsida Press, 2020.

[21] C. H. Lin and U. Nuha, "Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy," *J. Big Data*, vol. 10, no. 1, Dec. 2023, doi: 10.1186/s40537-023-00782-9.

[22] Juciananda Febriamita, Eliza Abelia, Nayla Zahratul Maula, and Ita Ita, "Pemerolehan Leksikon Ragam Bahasa Gaul pada Aplikasi X," *J. Pendidikan, Bhs. dan Budaya*, vol. 3, no. 4, pp. 62–69, 2024, doi: 10.55606/jpbb.v3i4.4560.

[23] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, Jul. 2020, doi: 10.5120/ijca2018917395.

[24] Aman Kharwal, "Classification Report in Machine Learning." Accessed: Nov. 18, 2024. [Online]. Available: https:// thecleverprogrammer .com /2021/07/07/classification-report-in-machine-learning/#google_vignette

[25] M. Bramer, *Principles of Data Mining*. London: Springer London, 2007. doi: 10.1007/978-1-84628-766-4.

[26] J. Winahyu and I. Suharjo, "Aplikasi Web Analisis Sentimen Dengan Algoritma Multinomial Naïve Bayes," *Kumpul. Artik. Mhs. Pendidik. Tek. Inform.*, vol. 10, no. 2, 2021.

[27] P. Naik, *Conceptualizing Python in Google COLAB*. India: Shashwat Publication, 2022. [Online]. Available: https:// www .researchgate .net/ publication/357929808

[28] F. A. Suharno and L. Listiyoko, "Aplikasi Berbasis Web dengan Metode Crawling sebagai Cara Pengumpulan Data untuk Mengambil Keputusan," in *Seminar Nasional Rekayasa Teknologi Informasi*, Tangerang, 2018, pp. 105–109.

[29] E. Nofiyanti and E. M. Oki Nur Haryanto, "Analisis Sentimen terhadap Penanggulangan Bencana di Indonesia," *J. Ilm. SINUS*, vol. 19, no. 2, p. 17, Jul. 2021, doi: 10.30646/sinus.v19i2.563.

[30] S. S. Mukrimaa *et al.*, "Machine Learning Teori, Studi Kasus dan Implementasi Menggunakan Python," 2016, *UR PRESS, RIAU*. doi: 10. 5281/zenodo.5113507.

[31] J. Yao, "Automated Sentiment Analysis of Text Data with NLTK," in *Journal of Physics: Conference Series*, China: Institute of Physics Publishing, May 2019. doi: 10.1088/1742-6596/1187/5/052020.

[32] S. A.Putra and A. Wijaya, "Analisis Sentimen *Artificial Intelligence* (AI) Pada Media Sosial Twitter Menggunakan Metode *Lexicon Based*," *J. Sist. Teknol. Inf. Komun. Vol.*, vol. 7, p. 23, 2023, [Online]. Available: Analisis Sentimen *Artificial Intelligence* (AI) Pada Media Sosial Twitter Menggunakan Metode *Lexicon Based*

[33] K. I. Gunawan and J. Santoso, "Multilabel *Text Classification* Menggunakan SVM dan Doc2Vec *Classification* Pada Dokumen Berita Bahasa Indonesia," *J. Inf. Syst. Hosp. Technol.*, vol. 3, no. 01, pp. 29–38, Apr. 2021, doi: 10.37823/insight.v3i01.126.

[34] M. R. Fais Sya' bani, U. Enri, and T. N. Padilah, "Analisis Sentimen Terhadap Bakal Calon Presiden 2024 Dengan Algoritme Naïve Bayes," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 2, p. 265, Apr. 2022, doi: 10.30865/jurikom.v9i2.3989.

[35] N. Husin, "Komparasi Algoritma Random Forest, Naïve Bayes, dan Bert Untuk Multi-Class Classification Pada Artikel Cable News Network (CNN)," *J. Esensi Infokom J. Esensi Sist. Inf. dan Sist. Komput.*, vol. 7, no. 1, pp. 75–84, 2023, doi: 10.55886/infokom.v7i1.608.

[36] H. Syuhada, "Analisis Sentimen Mengenai Komisi Pemberantasan Korupsi (KPK) Pada Media Sosial Twitter Dengan Menerapkan Algoritma Naive Bayes Classifier," Universitas Lampung, 2022.

[37] N. Attamami, A. Triayudi, and R. T. Aldisa, "Analisis Performa Algoritma Klasifikasi Naive Bayes dan C4.5 untuk Prediksi Penerima Bantuan Jaminan Kesehatan," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 7, no. 2, pp. 262–269, 2023, doi: 10.35870/jtik.v7i2.756.

[38] D. F. Sjoraida, B. W. K. Guna, and D. Yudhakusuma, "Analisis Sentimen Film Dirty Vote Menggunakan BERT (B*idirectional Encoder Representations from Transformers*)," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 8, no. 2, pp. 393–404, 2024, doi: 10.35870/jtik. v8i2. 1580.

[39] N. Sholihah, F. F. Abdulloh, and M. Rahardi, "*Sentiment Analysis* on KPU *Performance* Post-2024 Election via YouTube *Comments Using* BERT," *sinkron*, vol. 8, no. 4, pp. 2222–2232, 2024, doi: 10.33395/ sinkron. v8i4.14040.

[40] M. D. Maulidan, S. Sumarlinda, and Sopingi, "Development of Sentiment Analysis System of Simple Pol Application on Google Play Store Using Naive Bayes Classifier Method and BERT Prediction," *J. dinda*, vol. 4, no. 2, pp. 65–70, 2024.

[41] D. S. Arum, S. Butsianto, and R. Astuti, "Analisis Sentimen Masyarakat Indonesia Terhadap Sea Games 2023 Di Twitter Dengan Metode Naïve Bayes," *J. Inf. Syst. Applied, Manag. Account. Res.*, vol. 7, no. 3, pp. 728–738, 2023, doi: 10.52362/jisamar.v7i3.1150.

[42] A. S. Sulaeman, A. Sujjada, and I. L. Kharisma, "Penerapan Algoritma Cerdas Bidirectional Encoder Refresentations From Transformers Dalam Menganalisis Opini Publik Terhadap Produk Yang Mengalami Boikot," *INOVTEK Polbeng - Seri Inform.*, vol. 9, no. 1, pp. 460–473, 2024, doi: 10.35314/isi.v9i1.4252.

[43] P. A. Riyantoko, T. M. Fahrudin, D. A. Prasetya, T. Trimono, and T. D. Timur, "Analisis Sentimen Sederhana Menggunakan Algoritma LSTM dan BERT untuk Klasifikasi Data Spam dan Non-Spam," *Pros. Semin. Nas. Sains Data*, vol. 2, no. 1, pp. 103–111, 2022, doi: 10.33005/senada.v2i1.53.

[44] A. Fauzi, A. H. Yunial, D. E. Saputro, and R. Saputra, "Optimalisasi Random Forest untuk Sentimen Bahasa Indonesia dengan GridSearch dan SMOTE," *J. Ilmu Komput. dan Sist. Inf.*, vol. 4, no. 2, pp. 202–217, 2025, doi: 10.70340/jirsi.v4i2.207.