ANALISIS *FAIRNESS* MODEL KLASIFIKASI SENTIMEN PUBLIK TERHADAP ISU REVISI UU TNI MENGGUNAKAN *INDOBERT* BERDASARKAN *DEMOGRAPHIC PARITY* PADA KELOMPOK *GENDER*

LAPORAN TUGAS AKHIR

Laporan ini Disusun untuk Memenuhi Salah Satu Syarat Memperoleh Gelar Sarjana Strata 1 (S1) pada Program Studi Teknik Informatika Fakultas Teknologi Industri Universitas Islam Sultan Agung Semarang

Disusun Oleh :

Zulham Prabandanu

32602100126

FAKULTAS TEKNOLOGI INDUSTRI

UNIVERSITAS ISLAM SULTAN AGUNG

SEMARANG

2025

FAIRNESS ANALYSIS OF THE PUBLIC SENTIMENT CLASSIFICATION MODEL ON THE ISSUE OF REVISING TNI LAW USING INDOBERT BASED ON DEMOGRAPHIC PARITY IN GENDER GROUPS

FINAL PROJECT

Proposed to complete the requirement to obtain a bachelor's degree (S1) at Informatics Engineering Departement of Industrial Technology Faculty Sultan Agung Islamic University

ARRANGED BY:

Zulham Prabandanu

32602100126

MAJORING OF INFORMATICS ENGINEERING

INDUSTRIAL TECHNOLOGY FACULTY S

ULTAN AGUNG ISLAMIC UNIVERSITY

SEMARANG

2025

LEMBAR PENGESAHAN TUGAS AKHIR

ANALISIS *FAIRNESS* MODEL KLASIFIKASI SENTIMEN PUBLIK TERHADAP ISU REVISI UU TNI MENGGUNAKAN *INDOBERT* BERDASARKAN *DEMOGRAPHIC PARITY* PADA KELOMPOK *GENDER*

ZULHAM PRABANDANU
NIM 32602100126

Telah dipertahankan di depan tim penguji ujian sarjana tugas akhir
Program Studi Teknik Informatika
Universitas Islam Sultan Agung
Pada tanggal : 26 - 11 - 2025

TIM PENGUJI UJIAN SARJANA :

Bagus Satrio Waluyo Poetro,S.Kom.,M.Cs NIK.210616051 (Ketua Penguji)	.....	05/11/2025
Imam Much Ibnu Soebroto,S.T M.Sc P.hD NIK.210600017 (Anggota Penguji)	.....	05/11-2025
Badie'ah,S.T M.Kom NIK.210615044 (Pembimbing)	.....	26/11/2025

Semarang, 26 - 11 - 2025
Mengetahui,
Kaprodi Teknik Informatika
Universitas Islam Sultan Agung

Moch. Taufik, ST.,MIT.
NIK. 210604034

iii

SURAT PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan dibawah ini :

Nama : Zulham Prabandanu

NIM : 32602100126

Judul Tugas Akhir : ANALISIS *FAIRNESS* MODEL KLASIFIKASI SENTIMEN PUBLIK TERHADAP ISU REVISI UU TNI MENGGUNAKAN *INDOBERT* BERDASARKAN *DEMOGRAPHIC PARITY* PADA KELOMPOK *GENDER*.

Dengan bahwa ini saya menyatakan bahwa judul dan isi Tugas Akhir yang saya buat dalam rangka menyelesaikan Pendidikan Strata Satu (S1) Teknik Informatika tersebut adalah asli dan belum pernah diangkat, ditulis ataupun dipublikasikan oleh siapapun baik keseluruhan maupun sebagian, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka, dan apbila di kemudian hari ternyata terbukti bahwa judul Tugas Akhir tersebut pernah diangkat, ditulis ataupun dipublikasikan, maka saya bersedia dikenakan sanksi akademis. Demikian surat pernyataan ini saya buat dengan sadar dan penuh tanggung jawab.

Semarang, 3 - 12 - 2025

Yang Menyatakan,

Zulham Prabandanu

iv

PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

Saya yang bertanda tangan dibawah ini :

Nama : Zulham Prabandanu

NIM : 32602100126

Program Studi : Teknik Informatika

Fakultas : Teknologi industri

Dengan ini menyatakan Karya Ilmiah berupa Tugas akhir dengan Judul : **ANALISIS *FAIRNESS* MODEL KLASIFIKASI SENTIMEN PUBLIK TERHADAP ISU REVISI UU TNI MENGGUNAKAN *INDOBERT* BERDASARKAN *DEMOGRAPHIC PARITY* PADA KELOMPOK *GENDER*.**

Menyetujui menjadi hak milik Universitas Islam Sultan Agung serta memberikan Hak bebas Royalti Non-Eksklusif untuk disimpan, dialihmediakan, dikelola dan pangkalan data dan dipublikasikan diinternet dan media lain untuk kepentingan akademis selama tetap menyantumkan nama penulis sebagai pemilik hak cipta. Pernyataan ini saya buat dengan sungguh-sungguh. Apabila dikemudian hari terbukti ada pelanggaran Hak Cipta/Plagiarisme dalam karya ilmiah ini, maka segala bentuk tuntutan hukum yang timbul akan saya tanggung secara pribadi tanpa melibatkan Universitas Islam Sultan agung.

Semarang, 3-12-2025

Yang Menyatakan

Zulham Prabandanu

KATA PENGANTAR

Puji syukur Alhamdulillah atas kehadirat Allah SWT karena berkat rahmat dan hidayahnya penulis dapat menyelesaikan penulisan skripsi dengan judul "Analisis *Fairness* Model Klasifikasi *Sentiment Public* Terhadap Isu Revisi UU Tni Menggunakan *Indobert* Berdasarkan *Demographic Parity* Pada Kelompok *Gender*" untuk memenuhi salah satu sayarat menyelesaikan studi serta memperoleh gelar sarjana (S-1) pada program studi Teknik Informatika Fakultas Teknologi Industri Universitas Islam Sultan Agung Semarang.

Penelitian ini dibuat dan disusun dengan adanya bantuan dari berbagai pihak baik materi maupun teknis, oleh karena itu saya selaku penulis mengucapkan banyak terimakasih kepada:

1. Rektor UNISSULA Bapak Prof. Dr. H. Gunarto, S.H., M.H yang mengizinkan penulis menimba ilmu di kampus ini.
2. Dekan Fakultas Teknologi Industri Ibu Dr. Novi Marlyana, S.T., M.T.
3. Dosen pembimbing I ibu Badie'ah S.T., M.Kom yang telah meluangkan waktu dan memberi ilmu. Serta memberikan banyak nasehat dan saran.
4. Orang tua penulis yang telah mengizinkan untuk menyelesaikan laporan ini,
5. Dan kepada semua pihak yang tidak dapat saya sebutkan satu per satu

Dengan segala hormat saya, penulis masih menyadari bahwa masih banyak kekurangan dari segi kualitas maupun dari ilmu pengetahuan dalam menyusun laporan, sehingga penulis mengharapkan saran maupun kritikan membangun demi kesempurnaan laporan ini

Semarang, 5-12-2025

Zulham Prabandanu

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Media sosial menjadi sarana utama masyrakat dalam menyalurkan aspirasi pikiran terhadap kritik kebijakan pemerintah. Tagar #TolakRevisiUUTNI merupakan sebuah ungkapan kekecewaan masyarakat terhadap revisi yang menuai kontroversi. penelitian ini bertujuan untuk mengklasifikasi sentimen dan mengkaji hubungan *gender* dan kebijakan tersebut berdasarkan perbandingan beberapa model dan melakukan evaluasi *fairness* serta menerapkan mitigasi terhadap *bias* yang muncul. *IndoBERT (indobenchmark/indobert-base-p1)* yang akan di terapkan dalam penelitian ini, dengan parameter pelatihan 3 *epoch,* 16 *batch size* dan *learning rate* 2e-5. dataset yang sudah di labeli manual oleh dua annotator dan melakukan evaluasi terhadap hasil pelatihan pada model model dan juga mitigasi terhadap *bias* dengan salah satu teknik mitigasi *fairness* yaitu *Demographic Parity.* Hasil penelitian menunjukkan bahwa Model 3, dengan hasil terbaik dalam performa dan juga trade off. Dari segi keadilan, Model 3 memiliki nilai Demographic Parity Difference (DPD) sebesar 0.000 dan Demographic Parity Ratio (DPR) sebesar 1.000, yang menunjukkan tingkat keadilan yang baik tanpa perbedaan signifikan antar kelompok gender. Studi ini menunjukkan bahwa distribusi data yang seimbang dan penggunaan metode mitigasi fairness sangat penting untuk mencegah bias dalam sistem analisis sentimen berbasis pembelajaran mesin. Hasil ini membantu mengembangkan model AI yang tidak hanya akurat, tetapi juga adil dan inklusif.

Kata Kunci : Analisis Sentimen, *IndoBERT, Gender, Fairness, Demographic Parity*

ABSTRACT

Social media has become a primary medium for the public to express aspirations and criticism toward government policies. The hashtag #TolakRevisiUUTNI represents public dissatisfaction with a controversial legislative revision. This study aims to perform sentiment classification and examine its relationship with gender attributes while also evaluating fairness aspects in the applied models. The IndoBERT (indobenchmark/indobert-base-p1) model was implemented with training parameters of 3 epochs, a batch size of 16, and a learning rate of 2e-5. The dataset was manually labeled by two annotators, followed by model performance evaluation and bias mitigation using the Demographic Parity approach. The findings demonstrate that Model 3, which had a performed and trade off. The Demographic Parity Ratio (DPR) of 1.012 and Demographic Parity Difference (DPD) of 0.005 show that there are no appreciable discrepancies across gender groupings, indicating a strong degree of fairness. Model 3 showed the best balance between fairness and performance when compared to other models. This study emphasizes how crucial it is to distribute data in a balanced manner and use fairness mitigation strategies to avoid bias in machine learning-based sentiment analysis systems. The results aid in the creation of artificial intelligence models that are inclusive, equitable, and accurate.

Keyowrds: Sentiment Analysis, IndoBERT, Gender, Fairness, Demographic Parity

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pemerintah merupakan sekelompok atau sekumpulan orang dalam sebuah negara yang memiliki peran penting dalam penggerakan dan mengatur segala sesuatu yang bertujuan untuk memajukan dan menjalankan sebuah negara. Tanpa adanya pemerintahan yang ada sebuah negara tidak dapat berjalan dengan semestinya. Konsep ini terkait dengan bagaimana pemerintah menjalankan tugasnya secara efektif dan memberikan pelayanan publik yang berkualitas, transparan, akuntabel, dan juga adil(Resmadiktia dkk, 2023). Dengan demikian pemirintah memiliki peran yang signifikan dalam roda hidup sebuah negara.

Media sosial menjadi suatu jembatan aspirasi paling mudah bagi masyarakat agar suara ataupun keluhan mereka dapat di dengar oleh pemerintah. Media Sosial (*Social Media*) adalah saluran atau sarana pergaulan sosial secara *online* di dunia maya (*internet*)(Siregar, 2022). Salah satu media sosial yang popular dan diminati oleh banyak kalangan manusia di segala umur yaitu Twitter(X). penggunaan media sosial dapat memiliki fungsi kritis dalam hal mempertukarkan wacana, membangun kesadaran, hingga menciptakan inovasi(Alkatiri dkk, 2020). Twitter sebagai jembatan bagi semua masyarakat dalam memberikan aspirasi mereka terhdapa kinerja dan penetapan keputusan yang dilakukan pemerintah, sehingga tidak banyak kritikan maupun saran yang membangun yang di sampaikan oleh masyarakat luas. Namun akhir akhir ini banyak kebijakan pemerintah yang dianggap ngawur dan merugikan masyarakat banyak terutama bagi masyarakat menengah ke bawah sehingga menimbulkan sebuah topik baru yaitu # TolakRevisiUUTNI dimana tagar tersebut digunakan guna melakukan kritikan terhadap pemerintah yang melakukan kebijakan kebijakan yang kurang menguntungkan masyarakat.

Tagar #TolakRevisiUUTNI sebuah ungkapan ketidakpuasan masyarakat terhadap kebijakan tersebut, terlebih lagi dengan adanya presiden yang terlantik membuat sebuah kegaduhan dengan statemen dan upaya kebijakan yang akan di

laksanakannya. Maraknya kasus KKN yang mencuat juga menjadi alasan yang kuat adanya tren tagar tersebut dalam Twitter. Terhitung sejak awal tagar itu terbuat sudah lebih dari 350 ribu postingan yang menggunakan tagar tersebut sebagai bentuk protes masyarakat terhadap kebijakan yang akan di tetapkan. Pemilihan tagar tersebut merupakan sebuah studi kasus dalam penelitian yang di dasarkan pada tingginya respons masyrakat luas terhadap luasnya persebaran isu dan kritikan terhadap aturan dan kebijakan pemerintah yang mewakili opini masyarakat luas.

BERT (*Bidirectional Encoder Representations from Transformers*) merupakan sebuah model *machine learning* yang dilatih dengan data yang cukup banyak. Algoritma ini adalah invasi dari model *Transformer* dimana model tersebut memproses sebuah kata pada kalimat berdasarkan ada atau tidaknya kaitan antara kata tersebut dengan kalimat secara keseluruhan(Nayla dkk 2023). *IndoBERT* merupakan rumpun yang sama seperti *BERT* dimana dalam *IndoBERT* di spesifikasikan untuk mengolah data berbahasa Indonesia dengan *dataset* yang tersedia. Dengan digunakannya *IndoBERT* ini diharapkan dapat memproses dataset berbahasa Indonesia dengan cukup baik dan akurat.

Namun, masalah bias sering muncul saat menggunakan model kecerdasan buatan seperti IndoBERT, yang dapat memengaruhi keadilan hasil klasifikasi. Bias ini dapat berasal dari distribusi data yang tidak seimbang antar kelompok gender. Akibatnya, kelompok tertentu dapat menerima hasil yang tidak adil dari model. Oleh karena itu, penelitian ini juga melihat aspek keadilan model dengan menggunakan metrik kesetaraan demografis. Metrik ini digunakan untuk mengukur sejauh mana model memberikan peluang prediksi yang setara untuk masing-masing kelompok gender. Dengan cara ini, diharapkan bahwa hasil klasifikasi akan akurat dan adil secara demografis.

Lalu untuk pengertian bias sendiri yaitu suatu kecenderungan sebuah sistem atau model dalam memberikan hasil yang tidak adil atau kurang akurat terhadap suatu kelompok tertentu. digunakan untuk menggambarkan berbagai macam perilaku sistem, meskipun mereka mungkin berbahaya dengan cara yang berbeda, untuk kelompok yang berbeda, atau untuk alasan yang berbeda(Czarnowska dkk, 2021). Sedangkan untuk fairness yaitu memastikan model atau sistem tersebut tidak

condong pada suatu individu ataupun kelompok. *Fairness* merupakan keadaan dimana suatu data atau keputusan dianggap adil dan tidak terjadi diskriminasi baik antar individu atau kelompok(Wardani dkk, 2023). Dengan kata lain antara bias dan fairness ini memiliki keterikatan satu sama lain, jadi apabila penanganan bias tidak tercapai maka tidak akan mewujudkan suatu model yang adil atau fair.

Sementara itu dalam mengukur suatu model yang adil dan fair dengan mengukur menggunakan sebuah metrik dari demographic parity. Untuk menilai keadilan model, penelitian ini menggunakan metrik Demographic Parity, yang mengukur sejauh mana model memberikan hasil prediksi yang setara bagi setiap kelompok gender tanpa adanya kecenderungan pada salah satu pihak. Demografi Parity adalah metrik keadilan yang bertujuan untuk memastikan bahwa kemungkinan menerima hasil positif, seperti persetujuan pinjaman atau tawaran pekerjaan, adalah sama di seluruh kelompok yang didefinisikan oleh atribut sensitif seperti gender, ras, atau usia. Dengan kata lain, Demografi Parity terjadi ketika prediksi model tidak bergantung pada siapa yang termasuk dalam kelompok sensitif tersebut(Zeng, Dobriban dan Cheng, 2022).

Penelitian ini tidak hanya berfokus dalam klasifikasi dari sentiment tetapi mengkaji mengenai hubungan *gender* dengan kebijakan pemerintah yang di terapkan, hal ini menjadi penting untuk melihat apakah ada *bias* dalam penerimaan atau penolakan yang terjadi dalam masyarakat terutama dalam aspek *gender*. Dengan dilakukannya penelitian dengan menerapkan metode *IndoBERT* dalam mengamati hal-hal yang terjadi dalam tagar #TolakRevisiUUTNI dengan mengklasifikasikan sentiment dan distribusi dari *gender* yang akan di klasifikasikan menggunakan pendekatan *indoBERT* sebagai metode yang digunakan.

1.2 Rumusan Masalah

1. Bagaimana performa model klasifikasi sentiment publik terhadap revisi UU TNI berdasarkan akurasi dan distribusi prediksi antar *gender* (Pria dan Wanita)
2. Bagaimana *Fairness* dapat meningkatkan *demographic parity* pada model klasifikasi sentimen?

3. Bagaimana perbandingan akurasi dan *Fairness* metriks antara model sebelum dan sesudah penerapan teknik *Fairness*?

1.3 Batasan Masalah

1. Penelitian ini hanya berdasarkan tagar yang sedang tren akhir akhir ini dan tidak melakukan analisis diluar tagar tersebut dan hanya berfokus pada *platform* sosial media twitter.
2. Klasifikasi sentimen berdasarkan *gender* dilakukan dari analisis teks dan kemungkinan sebagai dukungan pelabelan secara manual.
3. Model yang digunakan dalam penelitian ini adalah *IndoBERT* sebagai representasi model bahasa berbasis *transformer*.

1.4 Tujuan Penelitian

1. Mengetahui Performa model klasifikasi sentimen dengan meninjau akurasi dan distribusi prediksi berdasarkan *gender*
2. Mengevaluasi dampak penerapan *fairness* dalam meningkatkan *Demographic Parity* dalam model.
3. Membandingkan performa model dari segi akurasi dan metrik *fairness* sebelum dan sesudah penerapan teknik *fairness*.

1.5 Manfaat

Dari penelitian ini diharapkan bahwa menunjukkan betapa pentingnya menggunakan prinsip keadilan dalam sistem klasifikasi sentimen berbasis *gender*. Penelitian ini difokuskan terutama pada cara analisis media sosial terhadap masalah kebijakan publik. Penelitian ini dapat memberikan wawasan tentang bagaimana model yang tampak lebih akurat belum tentu adil dalam memperlakukan kelompok *gender* yang berbeda dengan melihat dan membandingkan kinerja berbagai model klasifikasi dari segi akurasi dan metrik *fairness* seperti *demographic parity*. Studi ini diharapkan dapat berfungsi sebagai referensi untuk pengembangan sistem kecerdasan buatan yang lebih etis, inklusif, dan bertanggung jawab. Mereka juga

akan mendorong penggunaan model bahasa seperti *IndoBERT* untuk memahami opini publik secara menyeluruh tanpa bias terhadap kelompok tertentu.

1.6 Sistematika penulisan

Sistematika yang akan di gunakan oleh penulis sebagai berikut:

BAB I : PENDAHULUAN

BAB I Pendahuluan berisikan latar belakang, alasan pemilihan judul, tujuan, rumusan masalah, batasan masalah serta sistematika penulisan yang akan digunakan.

BAB II : TINJAUAN PUSTAKA DAN DASAR TEORI

BAB II yaitu tinjauan pustaka dan dasar teori, bab ini berisi mengenai teori teori yang akan digunakan dan memperkuat penelitian ini. Penelitian penelitian terdahulu juga menjadi alasan yang kuat dalam melakukan penelitian ini.

BAB III : METODE PENELITIAN

BAB III Menjelaskan mengenai langkah-langkah yang akan di tempuh dalam melakukan penelitian seperti pengumpulan *dataset*, tahap *preprocessing* data yang terdiri dari tokenisasi, *stemming*, *stopword removal* kemudian dilanjutkan tahapan *labelling* data, *training* model dan yang terakhir evaluasi model

BAB IV : HASIL DAN ANALISIS PENELITIAN

BAB IV menyajikan hasil yang diperoleh dari penelitian berdasarkan tahapan tahapan yang telah di lakukan. Berupa grafik hasil analisis yang di dapatkan berupa respon positif, negatif maupun netral berdasarkan respons masyarakat dalam media sosial.

BAB V : KESIMPULAN DAN SARAN

BAB V berisikan mengenai kesimpulan dari hasil penelitian tersebut dari awal dilakukannya penelitian hingga akhir. Seperti apa saja yang ditemukan ketika melakukan penelitian tersebut hingga paparan hasil yang di dapatkan. Serta kesimpulan untuk peneliti kedepannya jika akan melakukan pembaharuan terhadap penelitian tersebut

BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Dalam penelitian terdahulu penerapan model *IndoBERT* sudah dapat di implementasikan. Dalam sebuah penelitian Analisis Sentimen Pengguna X Indonesia Terkait Kendaraan Listrik dimana metode tersebut dapat di terapkan dengan menggunakan salah satu arsitekturnya yaitu *indoNLU* model yang dilatih dengan data *IndoNLU* memberikan kinerja yang lebih baik dalam memprediksi sentimen pada teks *tweet*. Evaluasi metrik seperti akurasi, presisi, *recall*, dan *F1-score* menunjukkan peningkatan yang signifikan pada model dengan data *IndoNLU*, serta lebih konsisten dalam setiap *epoch*(Merdiansah dan Ali Ridha, 2024).

Pada penelitian lain juga berhasil menerapkan metode *indoBERT* dengan penerapannya dalam Prediksi Emosi Dalam Teks Bahasa Indonesia. model ini dapat diandalkan untuk mendeteksi emosi tersebut. Selain itu, kelas "Senang" juga memiliki hasil yang cukup baik, meskipun terdapat beberapa kekeliruan dalam klasifikasi(Saputra dkk, 2025). Hal tersebut menyimpulkan bahwa metode ini menghasilkan hasil yang cukup baik dalam penggunaan metode *indoBERT* ini.

Penelitian terdahulu yang menggunakan metode *indoBERT* juga menghasilkan hasil yang cukup akurat dalam memahami data dalam Bahasa Indonesia dimana penggunaan pada Analisis Sentimen pada Pemilihan Presiden Indonesia. *IndoBERT* dapat di *fine-tuned* pada permasalahan *NLP* berbahasa Indonesia, memungkinkannya mampu memahami struktur bahasa, kosakata, dan konteks bahasa Indonesia lebih spesifik pada penelitian ini(Mahira Putri dkk, 2023).

Selain itu penelitian lain juga menerapkan metode *indoBERT* dalam melakukan sebuah penelitian Analisis Sentimen Terhadap Pembelajaran Secara Daring Pasca Pandemi *Covid-19*. klasifikasi opini menggunakan metode *IndoBERT* dapat membantu untuk mengklasifikasikan opini secara akurat dan membaginya menjadi 3 kategori (positif, netral dan negatif) secara otomatis(Hidayat dan Pramudita, 2024).

Seiring dengan berkembangnya model pembelajaran mesin isu mengenai keadilan menjadi sebuah sorotan penting guna membangun suatu model yang adil dan memastikan model tidak memihak suatu kelompok tertentu. mengharuskan kelompok minoritas (didefinisikan berdasarkan, misalnya, ras atau gender) untuk menerima hasil positif pada tingkat yang sama dengan kelompok mayoritas(Ashktorab dkk., 2023). Menghasilkan strategi yang digunakan praktisi pembelajaran mesin untuk menentukan apakah suatu model adil Kami juga mengidentifikasi berbagai strategi yang digunakan pengguna untuk menentukan kewajaran suatu model, membentuk kelompok identitas dari kategori serupa atau kelompok identitas dari kategori berbeda (ras, seksualitas, gender)(Ashktorab dkk., 2023)

Demographic Parity merupakan sebuah metode dalam menangani dan mengevaluasi sebuah model yang memiliki potensi *bias* untuk mewujudkan sebuah model yang adil. kekhawatiran bahwa model klasifikasi dokumen dapat bersifat diskriminatif dan dapat melestarikan bias manusia(Huang, 2022). Dari penelitian tersebut menyimpulkan bahwa Eksperimen menunjukkan bahwa dengan memperlakukan kelompok demografis sebagai domain, kami dapat mengurangi bias sekaligus mempertahankan kinerja yang relatif baik(Huang, 2022).

Isu mengenai sebuah revisi undang undang militer menjadi sebuah pemicu perdebatan terhadap peran militer dalam ranah public memunculkan kekhawatiran masyarakat luas. sentimen media online mayoritas positif (74%) dan mendukung revisi dengan narasi bahwa RUU ini diperlukan untuk kebutuhan strategis, serta menegaskan bahwa prajurit aktif yang menduduki jabatan sipil harus mengundurkan diri. Namun, 13% mencatatkan sentimen negatif, terutama menyoroti isu transparansi dan lokasi rapat yang tertutup(Utami, 2025). Sebaliknya, sentimen di media sosial sangat dominan negatif (81%). Publik mempertanyakan transparansi proses legislasi, mengkhawatirkan kembalinya dwifungsi TNI, serta menilai revisi ini sebagai kemunduran dari semangat reformasi militer. Hanya 14% akun yang mendukung revisi(Utami, 2025).

Dari penelitian terdahulu memunculkan sebuah kekurangan dimana dari penelitian analisis sentiment terutama dalam konteks Bahasa Indonesia masih

belum membahas mengenai sebuah keadilan dalam model yang di bangun, hal ini menjadi sebuah dorongan untuk meneliti dan membangun sebuah model dengan mengedepankan aspek keadilan dalam model sehingga tidak menimbulkan *bias* dalam model tersebut. Berdasarkan tinjauan pustaka tersebut dapat diambil kesimpulan bahwa dengan penggunaan *indoBERT* menghasilkan hasil yang memuaskan dan positif terhadap hasil yang diharapkan, dengan demikian penelitian ini akan menggunakan metode *indoBERT* dalam peenggunaan penelitian kedepannya. Selain itu berdasarkan dari topik yang diangkat memunculkan sebuah permasalahan *bias* yang dapat di teliti yaitu *Gender* dengan demikian dilakukannya penelitian ini akan memanfaatkan metode *indoBERT* dalam melakukan klasifikasi *gender* dan sentiment berdasarkan teks pada dataset diamana keunggulan *indoBERT* yang dapat memahami teks terutama ber Bahasa Indonesia.

Tabel 2. 1 Tinjauan Pustaka

Penulis	**Judul**	**Hasil Penelitian**
Ade Chandra Saputra, Agus Sehatman Saragih , Deddy Ronaldo (2025)	Prediksi Emosi Dalam Teks Bahasa Indonesia Menggunakan Model *Indobert*	model ini mampu mengklasifikasikan enam emosi dengan akurasi keseluruhan sebesar 73%. Model menunjukkan performa terbaik dalam mengklasifikasikan emosi "Jijik" dengan tingkat precision dan recall yang tinggi, yaitu 0.98, menunjukkan bahwa model ini dapat diandalkan untuk mendeteksi emosi tersebut.
Dharmawan, Steven Mawardi, Viny Christanti Perdana, Novario Jaya(2023)	Klasifikasi Ujaran Kebencian Menggunakan Metode *FeedForward Neural*	Metode *feedforward neural network* dengan *IndoBERT* berhasil melakukan klasifikasi dengan nilai akurasi terbaik sebesar 89,52%.

Penulis	Judul	Hasil Penelitian
	Network (IndoBERT)	
Hidayat, Muhammad Nur Pramudita, Rully (2024)	Analisis Sentimen Terhadap Pembelajaran Secara Daring Pasca Pandemi Covid-19 Menggunakan Metode *IndoBERT*	sistem klasifikasi opini menggunakan metode *IndoBERT* dapat membantu untuk mengklasifikasikan opini secara akurat dan membaginya menjadi 3 kategori (positif, netral dan negatif) secara otomatis dengan nilai akurasi sebesar 0,87 atau 87%. Penelitian
Imron, Syaiful Setiawan, Esther Irawati Santoso, Joan. (2023)	Deteksi Aspek Review *E-Commerce* Menggunakan *IndoBERT Embedding* dan *CNN*	Setelah dilakukan pengujian dan evaluasi, penelitian ekstraksi aspek dengan menggunakan *BERT* sebagai *word embedding* dan metode *CNN* untuk ekstraksi aspek mendapatkan hasil yang sangat baik, yaitu akurasi sebesar 94,86%.

2.2 Dasar Teori

2.2.1 Pemerintah

pemerintah adalah sekumpulan khusus dari individu-individu yang telah menetapkan tanggungjawab untuk mempertahankan dan/atau mengadaptasi system dimana mereka menjadi bagiannya. Menjalankan tanggung jawab dengan membuat pilihan-pilihan yang mengikat para anggotanya(Makalew, 2021). Sedangkan menurut peneliti yang lain, Pemerintahan secara etimologi kata pemerintah berasal

dari kata “perintah"yang berarti sesuatu yang harus dilaksanakan,yang kemudian mendapat imbuhan sebagai berikut(La Dahiri, 2020).

2.2.2 Media Sosial dan Twitter

Twitter merupakan bagian dari media social yang di kelola oleh *Elon Musk* yang mendukung orang orang untuk melakukan kebebasan ber ekspresi secara terbuka baik dari lapisan masyarakat hingga ke lapisan pemerintah. Melalui *platform* tersebut mempermudah melakukan jembatan aspirasi antara masyarakat dengan lembaga pemerintahan. Salah satu media sosial yang sering digunakan untuk mempertukarkan wacana adalah Twitter. Twitter dapat mendorong sentimen publik dan mengatur kemarahan publik, simpati, sukacita, dan ketakutan(Alkatiri dkk, 2020).

Twitter menjadi sebuah media acuan yang cukup bagus dalam melakukan sebuah penelitian terhadap isu yang sedang hangat, ditambah twitter juga menjadi sebuah media yang *up to date* dalam mengamati perkembangan isu yang ada.

Model *indoBERT* sudah banyak di terapkan dalam melakukan analisis pada dataset yang di peroleh melalui twitter, Berdasarkan hasil evaluasi, model *Indo-BERT* memiliki potensi besar untuk mendsukung berbagai aplikasi, seperti pemantauan opini publik, analisis risiko, dan pemahaman konten media sosial(Nurjoko dan Agus Rahardi, 2024)

2.2.3 Analisis sentimen

Analisis sentimen adalah salah satu cabang dari Natural Language Processing (NLP) yang bertujuan untuk mengklasifikasikan opini dalam teks menjadi kategori seperti positif, negatif, atau netral(Andriyani dkk., 2025). Tujuan analisis sentimen yang bertujuan untuk mengidentifikasi opini berdasarkan data teks, memahami pandangan publik menjadi penting dalam membentuk sentimen publik dan memengaruhi pilihan(Sayarizki dan Nurrahmi, 2024).

Sementara itu metode metode dalam analisis sentimen meninjau teknik pemrosesan bahasa alami untuk mengekstraksi fitur berdasarkan jenis kata dan posisi istilah teknik statistik untuk mengekstraksi fitur berdasarkan frekuensi kata dan model pohon keputusan dan teknik untuk menggabungkan penandaan jenis kata, analisis fitur sintaksis, dan kamus(Cui dkk., 2023). fitur terbaik saat ini untuk

analisis sentimen teks Twitter adalah AFINN (daftar istilah bahasa Inggris yang digunakan untuk analisis sentimen yang dinilai secara manual oleh Finn Årup Nielsen(Cui dkk., 2023). metode pembelajaran mendalam yang digunakan dalam berbagai aplikasi pada tingkat analisis sentimen kalimat dan aspek/objek, termasuk Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), dan Long Short-term Memory (LSTM). kelebihan dan kekurangan metode-metode ini serta parameter kinerjanya, memperkenalkan teknik pembelajaran mendalam seperti Jaringan Saraf Tiruan Dalam (DNN), CNN, dan Jaringan Kepercayaan Dalam (DBN) untuk menyelesaikan tugas-tugas analisis sentimen seperti klasifikasi sentimen, masalah lintas bahasa dan analisis ulasan produk. menyelidiki teknik pembelajaran mendalam dan pembelajaran mesin untuk analisis sentimen dalam konteks ekstraksi dan kategorisasi aspek, ekstraksi ekspresi opini, ekstraksi pemegang opini, analisis sarkasme, data multimodal. membandingkan kinerja metode pembelajaran mendalam pada kumpulan data tertentu dan mengusulkan bahwa kinerja dapat ditingkatkan menggunakan model termasuk Representasi Encoder Dua Arah dari Transformer (BERT), model penyematan kata khusus sentimen, model perhatian berbasis kognitif, dan pengetahuan akal sehat.(Cui dkk., 2023)

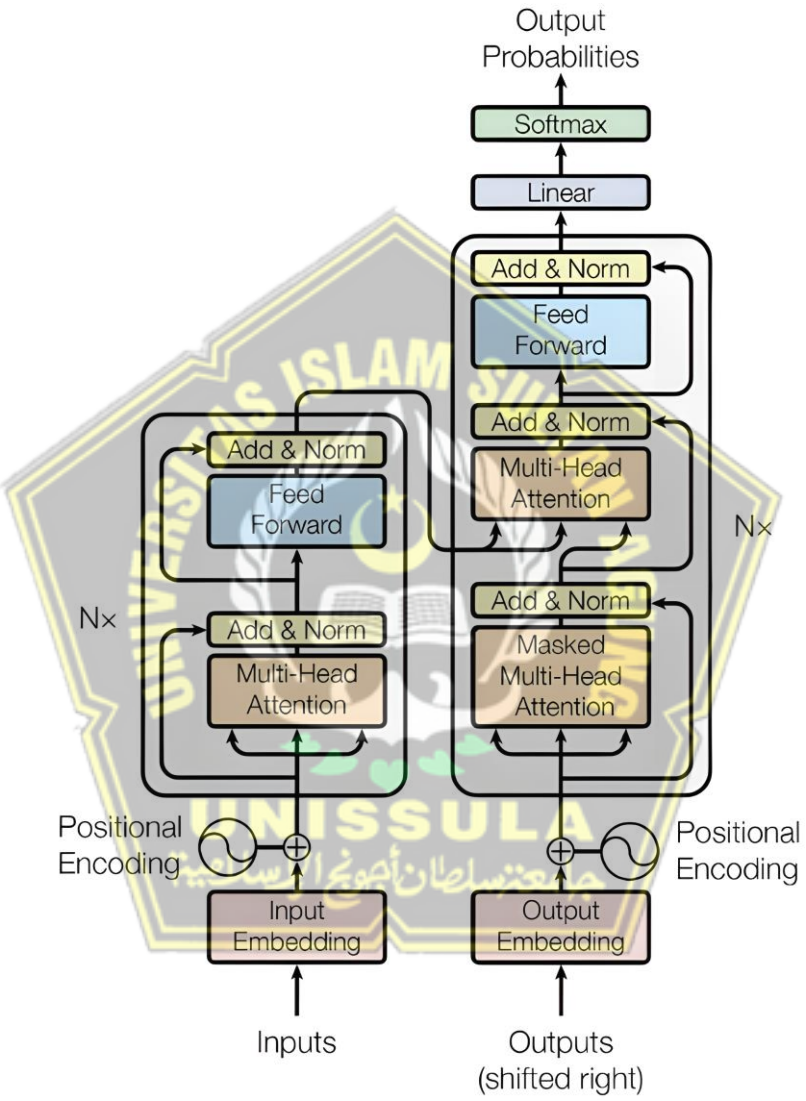
Selain itu hambatan ataupun tantangan dalam penggunaan Bahasa Indonesia yaitu perlu kajian lebih lanjut pada seluruh konjungsi dalam bahasa Indonesia agar pembagian klausa dalam kalimat sesuai dengan kaidah bahasa Indonesia(Saputra dkk., 2021).

2.2.4 Transformer

Transformer adalah serangkaian token. Keluaran dari encoder adalah representasi berdimensi tetap untuk setiap token beserta embedding terpisah untuk keseluruhan urutan tersebut(Amatriain dkk., 2024).

Encoder dan decoder adalah dua komponen utama dalam arsitektur transformer. Sementara encoder bertanggung jawab untuk memproses input dan membuat representasi kontekstual untuk setiap token, decoder menggunakan representasi tersebut untuk menghasilkan keluaran, seperti dalam tugas terjemahan mesin. Mekanisme utama yang digunakan adalah self-attention, yang

memungkinkan model untuk mengidentifikasi hubungan antar token tanpa memperhatikan jarak posisinya dalam urutan teks. Selain itu, Transformer menggunakan positional encoding untuk menambahkan informasi posisi ke dalam embedding token untuk menjaga urutan token tetap diperhatikan.



Gambar 2. 1 Arsitektur Transformer dari(Amatriain dkk., 2024)

Seluruh urutan input diproses oleh bagian encoder secara bersamaan. Dimulai dengan input embedding, proses ini mengubah kata-kata menjadi vektor numerik. Setelah itu, kata-kata ditambahkan dengan enkripsi posisi untuk menyimpan informasi posisinya. Selain itu, lapisan multi-head attention dan feed forward network digunakan untuk memproses data. Ini memungkinkan model untuk

memahami hubungan kontekstual antar kata dalam kalimat. Bagian decoder kemudian menerima output encoder(Mohiuddin dkk., 2023).

Decoder membuat urutan keluaran kata demi kata. Lapisan perhatian multi-kepala pada decoder terdiri dari dua kategori. Lapisan perhatian multi-kepala yang disembunyikan memastikan bahwa model hanya melihat kata-kata yang sudah diprediksi sebelumnya. Lapisan perhatian multi-kepala yang berinteraksi dengan output encoder untuk fokus pada bagian masukan yang relevan. Setelah itu, lapisan softmax dan linear digunakan untuk mengubah hasil decoder menjadi probabilitas kata-kata keluaran. Keunggulan utama Transformer adalah kemampuan untuk memproses data secara paralel, yang membuatnya jauh lebih efisien untuk urutan data yang panjang dibandingkan dengan model sebelumnya, seperti RNN(Mohiuddin dkk., 2023)

Sedangkan definisi *pre-training* adalah Kami menggambarkan arsitektur Transformer sebagai gabungan dari Encoder dan Decoder(Amatriain dkk., 2024). Sedangkan *fine tuning* mengacu pada penyempurnaan model dasar untuk tugas tertentu, seperti klasifikasi spam atau tanya jawab. Model, seperti BERT, menghasilkan representasi token input, tetapi tidak dapat menyelesaikan tugas apa pun dengan sendirinya. Oleh karena itu, perlu dilakukan fine-tuning dengan menambahkan lapisan neural tambahan di atas model dasar dan melatih model secara menyeluruh. Dengan demikian, Transformer memberikan fleksibilitas yang tinggi: pre-training memungkinkan model memahami bahasa secara umum, sementara fine-tuning mengadaptasikan pemahaman tersebut untuk menyelesaikan tugas spesifik(Amatriain dkk., 2024)

2.2.5 *IndoBERT*

BERT, atau Representations of Bidirectional Encoder from Transformers, adalah model bahasa inovatif yang dikembangkan oleh Google yang mengubah cara mesin memahami bahasa manusia. BERT mengidentifikasi hubungan kontekstual antar token dalam teks dalam dua arah, yaitu dari kiri ke kanan dan dari kanan ke kiri, melalui mekanisme self-attention. Model ini terdiri dari tumpukan (stack) transformer encoder yang masing-masing memiliki dua komponen utama:

neural network feed-forward yang dilengkapi dengan koneksi residual dan normalisasi lapisan(Devlin dkk., 2020).

Dua tujuan utama digunakan untuk melatih BERT sebelum pelatihan Model Bahasa Masked (MLM), di mana sejumlah token masukan diacak dan dimask untuk diprediksi kembali, dan Next Sentence Prediction (NSP), yang dimaksudkan untuk mengajarkan pemahaman hubungan antar kalimat. Setelah pelatihan sebelumnya, BERT dapat disesuaikan secara menyeluruh untuk melakukan berbagai tugas pengolahan bahasa natural, seperti mengklasifikasikan teks, menjawab pertanyaan, dan mengidentifikasi entitas bernama dengan menambahkan lapisan output sederhana di atas representasi yang dibuat. Karena kemampuan BERT untuk menangkap konteks bahasa yang kompleks secara bidirectional, pendekatan ini memungkinkan BERT untuk mencapai performa modern pada berbagai standar NLP(Devlin dkk., 2020).

IndoBERT adalah model bahasa berbasis arsitektur BERT yang telah dilatih khusus untuk korpus bahasa Indonesia yang luas, yang mencakup sekitar 5,5 miliar kata dalam berbagai jenis teks, seperti artikel berita, media sosial, dll.. Oleh karena itu, sangat cocok untuk digunakan untuk tugas-tugas yang melibatkan klasifikasi teks dan analisis sentimen dalam konteks Indonesia. Model ini memiliki struktur yang didasarkan pada BERT dan memiliki dua belas lapisan transformasi, 768 ukuran vektor tersembunyi (unit tersembunyi) dan dua belas kepala perhatian. Ini memungkinkan model untuk menangkap nuansa konteks dalam kalimat lebih dalam dan kompleks daripada metode konvensional seperti LSTM atau SVM(Cahyawijaya dkk., 2021).

IndoBERT merupakan sebuah metode yang khusus digunakan dalam proses training pada data ber-bahasa Indonesia, model ini telah dilatih dengan korpus 5,5 miliar kata yang mencakup beberapa bentuk teks bahasa Indonesia. Sehingga cocok di gunakan untuk melatih dataset yang akan digunakan nantinya, cara kerja metode ini yaitu *BERT* menggunakan mekanisme *self-attention*, yaitu menggabungkan beberapa vektor kata sebagai masukan dan menyertakan *self-attention* di kedua arah antara dua kalimat(Anugerah Simanjuntak*l.*, 2024)

IndoBERT akan melibatkan pengoptimalan untuk klasifikasi sentimen positif, netral, dan negatif dari teks Twitter berbahasa Indonesia. Diharapkan bahwa metode ini akan meningkatkan akurasi dan keadilan model karena, dalam memproses berbagai jenis teks media sosial, representasi token yang sadar konteks sangat penting. Studi empiris menunjukkan bahwa IndoBERT lebih baik daripada model klasik dalam berbagai tugas klasifikasi sentimen bahasa Indonesia(Dhendra dan Gayuh Utomo, 2025).

2.2.6 Keterkaitan *Gender* dengan keputusan Pemerintah

Gender adalah suatu konsep kultural yang dipakai untuk membedakan peran, perilaku, mentalitas, dan karakteristik emosional antara laki-laki dan perempuan yang berkembang dalam masyarakat(Nurhasanah dan Zuriatin, 2023). *Gender* lebih merujuk pada norma norma dan harapan sosial yang dapat mempengaruhi seorang individu dalam berinteraksi, berpartisipasi dalam ruang public termasuk dalam pemerintahan.

Isu *gender* memengaruhi representasi, partisipasi, dan pengaruh kebijakan terhadap kelompok masyarakat, yang membuatnya menjadi penting dalam konteks sosial-politik. *Gender* tidak hanya terbatas pada perempuan itu juga mencakup ketimpangan peran dan akses antar kelompok gender terhadap layanan publik, ruang demokrasi, dan pengambilan keputusan. hambatan terkait pemahaman perspektif *gender*, komunikasi yang kurang memadai antar organisasi, dan kurangnya upaya pemerintah dalam meningkatkan partisipasi masyarakat(Takayasa, 2023) hal tersebut akan menimbulkan ketimpangan yang terjadi dalam masyarakat terutama dalam *gender* dengan itu menunjukan bahwa pentingnya pemerintah dalam melakukan komunikasi dalam menerapkan kebijakan untuk memperhatikan semua kelompok. kesetaraan antara perempuan dan laki-laki atau kesetaraan *gender* mendorong partisipasi perempuan dan laki-laki dalam pengambilan keputusan mendukung perempuan dan anak perempuan sehingga mereka dapat sepenuhnya memperoleh hak-hak mereka dan mengurangi kesenjangan antara perempuan dan laki-laki dalam hal akses dan kontrol atas sumber daya dan manfaat dari pembangunan(Takayasa, 2023). Dengan demikian

dengan melihat *gender* dapat memutuskan sebuah keputusan atau kebijakan yang dapat bermanfaat bagi pembangunan dan keberlanjutan.

2.2.7 *Bias* dan *Fairness* pada model

Bias Merupakan suatu kecenderungan sebuah sistem atau model dalam memberikan hasil yang tidak adil atau kurang akurat terhadap suatu kelompok tertentu. digunakan untuk menggambarkan berbagai macam perilaku sistem, meskipun mereka mungkin berbahaya dengan cara yang berbeda, untuk kelompok yang berbeda, atau untuk alasan yang berbeda(Czarnowska dkk, 2021). Dengan kata lain bias berarti *bias* menghasilkan sebuah hasil yang tidak adil, basis untuk memilih suatu generalisasi (hipotesis) atas individu atau grup lain tanpa memperhatikan konsistensi yang ketat dengan pelatihan yang diamati(Wardani dkk, 2023). Dengan terdapatnya bias terdeput maka perlu diacapainya *Fairness.*

Fairness Yaitu memastikan model atau sistem tersebut tidak condong pada suatu individu ataupun kelompok. *Fairness* merupakan keadaan dimana suatu data atau keputusan dianggap adil dan tidak terjadi diskriminasi baik antar individu atau kelompok(Wardani dkk, 2023). Dengan menggunakan pendekatan *Demographic parity* dan *Equalized odds* kita dapat melakukan mitigasi pada ketidakadilan model tanpa mengubah atapun menambah data uji, Sebuah pengklasifikasi yang memenuhi paritas demografi di bawah distribusi (X,A,Y)jika prediksinyaH(X)secara statistik independen dari atribut yang dilindungi(Agarwal dkk., 2018). Sedangkan *equalized odds,* Sebuah pengklasifikasi yang memenuhi peluang yang sama di bawah distribusi atas(X,A,Y) jika prediksinyaH(X)secara kondisional independen dari atribut yang dilindungi(Agarwal dkk., 2018). *Demographic Parity* bertujuan untuk memastikan bahwa peluang setiap individu dalam kelompok sensitif untuk menerima prediksi positif adalah setara sedangkan *Equalized Odds* menekankan pada kesetaraan performa model terhadap setiap kelompok sensitif, dengan cara menyamakan *True Positive Rate (TPR)* dan *False Positive Rate (FPR)*

2.2.8 *Demographic Parity*

Dalam pembelajaran mesin, Demografi Parity adalah metrik keadilan yang bertujuan untuk memastikan bahwa kemungkinan menerima hasil positif, seperti

persetujuan pinjaman atau tawaran pekerjaan, adalah sama di seluruh kelompok yang didefinisikan oleh atribut sensitif seperti gender, ras, atau usia. Dengan kata lain, Demografi Parity terjadi ketika prediksi model tidak bergantung pada siapa yang termasuk dalam kelompok sensitif tersebut(Zeng, Dobriban dan Cheng, 2022).

Sementara itu jenis jenis bias yang di nilai untuk evaluasi mitigasi yaitu equal opportunity bias dimana ketika model pembelajaran mesin memberikan peluang yang tidak setara bagi individu yang memenuhi syarat untuk menerima hasil positif berdasarkan atribut sensitif seperti gender, ras, atau usia, terjadi bias peluang yang tidak setara. True Positive Rate (TPR) adalah metrik yang digunakan untuk mengukur bias ini. Jika TPR berbeda secara signifikan antara kelompok sensitif, maka model dianggap tidak adil menurut Equal Opportunity(Hardt, Price dan Srebro, 2020).

Selain itu, Ketika model menghasilkan tingkat false positive rate (FPR) yang berbeda pada kelompok sensitif, terjadi ketidakadilan prediktif dalam sistem pembelajaran mesin. Kondisi ini terjadi ketika suatu kelompok lebih sering dianggap sebagai kelas positif, meskipun sebenarnya termasuk dalam kelas negatif. Perbedaan FPR antara kelompok menunjukkan bahwa model tidak memperlakukan semua kelompok secara adil dalam hal kesalahan prediksi positif. Oleh karena itu, model dianggap memenuhi kriteria prediktor kesetaraan apabila nilai FPR pada setiap kelompok sensitif relatif sama, dan jika terjadi perbedaan nilai yang signifikan menunjukkan bias(Zeng, Dobriban dan Cheng, 2022).

Untuk memastikan bahwa model tidak merugikan kelompok tertentu, penelitian tentang keadilan dalam model machine learning secara fundamental bergantung pada metrik dasar. Salah satu metrik awal untuk mengukur keseimbangan demografi adalah pilihan rasio (SR), yang didefinisikan sebagai kondisi di mana model memberikan hasil positif pada tingkat yang sama di seluruh kelompok demografi. Karena akurasi model tidak diperhitungkan, SR saja tidak cukup. Oleh karena itu, metrik seperti Rate Positif Benar (TPR) dan Rate Negatif Positif (FPR) sangat penting. Perbedaan TPR yang signifikan antar kelompok

menunjukkan bias dalam memberikan kesempatan yang sama(Verma dan Rubin, 2020).

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

TP (True Positive) merupakan Jumlah hasil positif yang diprediksi dengan benar, sedangkan TN (True Negative) Jumlah hasil negatif yang diprediksi dengan benar. Lalu FP (False Positive) Jumlah hasil negatif yang diprediksi salah sebagai positif. Dan yang terakhir FN (False Negative) Jumlah hasil positif yang diprediksi salah sebagai negatif (Verma dan Rubin, 2020).

$$TPR = \frac{TP}{TP + FN} \quad (2.2)$$

TP (True Positive) Jumlah kasus positif yang diprediksi dengan benar. Lalu FN (False Negative) Jumlah kasus positif yang diprediksi salah sebagainegatif(Verma dan Rubin, 2020).

$$FPR = \frac{FP}{FP + TN} \quad (2.3)$$

FP (False Positive) Jumlah kasus negatif yang diprediksi salah sebagai positif. TN (True Negative) Jumlah kasus negatif yang diprediksi dengan benar (Verma dan Rubin, 2020).

$$SR = \frac{TP + FP}{TP + FP + FP + FN} \quad (2.4)$$

TP (True Positive) Jumlah kasus positif yang diprediksi dengan benar. FP (False Positive) Jumlah kasus negatif yang diprediksi salah sebagai positif. TN (True Negative) Jumlah kasus negatif yang diprediksi dengan benar. FN (False Negative) Jumlah kasus positif yang diprediksi salah sebagai negatif (Verma dan Rubin, 2020).

Sebaliknya, FPR mengukur predictive equality, yaitu seberapa sering model membuat kesalahan positif pada individu yang tidak berkualitas dari kelompok yang berbeda. Perbedaan FPR yang mencolok menunjukkan bahwa model lebih sering salah pada satu kelompok, membuatnya tidak adil. Selanjutnya, semua metrik ini dievaluasi bersama dengan akurasi model secara keseluruhan untuk

memastikan bahwa upaya mitigasi bias tidak berdampak negatif pada kinerja. Analisis menyeluruh ini adalah langkah penting menuju pembuatan sistem AI yang adil dan bertanggung jawab(Friedler dkk., 2020).

Sebuah pengklasifikasi memenuhi paritas demografis berdasarkan distribusi(Tang, Zhang dan Zhang, 2023). Persamaan yang digunakan dalam menghitung *demographic parity* menurut (Tang, Zhang dan Zhang, 2023):

$$\forall a, a' \in A : P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = a') \quad (2.5)$$

Rumus tersebut merupakan definisi formal dari demographic parity dalam fairness yang memiliki arti $\hat{Y}$ = prediksi model (1 = hasil positif), A = atribut sensitif (misal gender, ras, usia), a dan a' = dua nilai berbeda dari atribut sensitif.

Salah satu pendekatan yang paling umum digunakan dalam kajian fairness pada model machine learning adalah paritas demografis. Pendekatan ini menekankan bahwa keputusan model tidak boleh bergantung pada faktor sensitif seperti gender, ras, atau usia. Dengan kata lain, setiap kelompok populasi seharusnya memiliki peluang yang sama untuk mendapatkan hasil positif dari prediksi model, tidak peduli atribut sensitif apa pun yang dimilikinya (Verma dan Rubin, 2020).

$$DPD = P\ (\hat{Y} = 1 \mid A = a) - P\ (\hat{Y} = 1 \mid A = a') \quad (2.6)$$

$\hat{Y}$ untuk Prediksi hasil model, A Atribut sensitif (misalnya gender, ras, usia), a untuk Kelompok sensitif tertentu (misalnya "laki-laki"), a' untuk Kelompok sensitif pembanding (misalnya "perempuan"). Jika DPD = 0, maka model dianggap fair karena kedua kelompok mendapatkan tingkat prediksi positif yang sama. Jika DPD semakin besar (positif/negatif), berarti ada bias karena model lebih menguntungkan salah satu kelompok(Verma dan Rubin, 2020).

$$DPR = \frac{P\ (\hat{Y} = 1 \mid A = a)}{P\ (\hat{Y} = 1 \mid A = a')} \quad (2.7)$$

$\hat{Y}$ untuk Prediksi hasil model, A Atribut sensitif (misalnya gender, ras, usia), a a' untuk dua kelompok berbeda dalam atribut sensitif (contoh: laki-laki dan perempuan). DPR mengukur rasio probabilitas hasil positif antara dua kelompok, Jika DPR = 1, berarti kedua kelompok mendapat perlakuan setara, jika DPR < 1, kelompok minoritas lebih sedikit menerima hasil positif (Verma dan Rubin, 2020).

Salah satu metode post-processing dalam pembelajaran mesin fairness adalah Threshold Optimizer. Ini digunakan untuk menyesuaikan ambang keputusan (threshold) model prediksi untuk memenuhi beberapa kriteria fairness, salah satunya Demografi Parity. Model prediksi biasanya menghasilkan probabilitas atau skor kontinu, yang kemudian diubah menjadi prediksi biner, yang dapat positif atau negatif, dengan menggunakan ambang standar, seperti 0,5. Tetapi penetapan ambang yang sama untuk semua kelompok sensitif dapat menyebabkan bias demografis karena proporsi prediksi positif yang berbeda di antara kelompok tersebut. Dengan kata lain, Threshold Optimizer memaksimalkan keadilan berdasarkan paritas demografi tanpa mengubah model awal. Metode ini menggunakan pengurangan pasca-proses karena model tetap dipertahankan sementara prediksi akhir disesuaikan untuk menjadi adil(Agarwal dkk., 2020).

Dalam penelitian tersebut juga menyimpulkan mengenai kelebihan Keunggulannya termasuk kemampuan untuk mendorong representasi yang setara antar kelompok dengan berbagai karakteristik sensitif, konsepnya sederhana sehingga pemangku kepentingan non-teknis dapat memahaminya dengan mudah, dan cocok digunakan sebagai indikator awal untuk mengidentifikasi potensi bias sebelum menerapkan metrik yang lebih kompleks. Meskipun demikian, metrik ini memiliki beberapa masalah. Ini termasuk mengabaikan perbedaan distribusi dasar antar kelompok, yang dapat menyebabkan koreksi berlebihan, berpotensi menurunkan akurasi model karena memaksa perubahan prediksi yang benar menjadi salah, tidak menjamin keadilan pada tingkat individu, dan sulit dipenuhi bersamaan dengan metrik fairness lainnya seperti *Equalized Odds* atau *Predictive Parity* kecuali dalam kondisi distribusi tertentu(Tang, Zhang dan Zhang, 2023).

2.2.9 Anotasi Dataset *Gender* dan *Sentiment*

Labelling merupakan suatu upaya yang dilakukan untuk menandai sebuah data tertentu untuk tujuan penelitian dan lain lain. Pelabelan *gender* yang dilakukan dalam penelitian ini menggunakan cara manual yang dilakukan anatara 2 orang. Hal tersebut di dasari pada tata cara pelabelan dilakukan melalui anotasi manusia untuk memberi label pada tweet dengan label sentimen(Geni dkk, 2024). Lalu selanjutnya dilakukan persamaan dimana data yang memiliki nilai *sentiment* sama maka dapat

digunakan untuk proses lebih lanjut. Pada bagian *gender* juga dilakukan hal yang sama. Ia menganalisis bagaimana penggunaan bahasa bervariasi tergantung pada ciri-ciri pribadi(Rangel dkk., 2018). Untuk selanjutnya diberikan pelabelan secara manual, pengguna diberi label secara manual ke dalam tiga kelas, yaitu perempuan, laki-laki(Nia dkk., 2023) dari dataset yang sudah di kumpulkan, melatih model klasifikasi teks untuk pengenalan jenis kelamin dapat membantu mengekstrak jenis kelamin lebih banyak pengguna dan meningkatkan kinerja model(Nia dkk., 2023).

2.2.10 Evaluasi Model Klasifikasi Sentimen

Dalam evaluasi model yang dihasilkan lumrahnya menggunakan metriks metriks yang dihasilkan untuk mengukur seberapa baik model tersebut dalam belajar berdasarkan data uji .

$$accuracy = \frac{true\ positive + true\ negative}{total\ prediction} \tag{2.8}$$

Dalam akurasi mengukur keseluruhan proporsi prediksi yang benar, Meskipun akurasi merupakan metrik yang lugas, akurasi dapat menyesatkan dalam kumpulan data yang tidak seimbang(Kaur dan Kaur Sandhu, 2023).

$$precission = \frac{true\ positive}{true\ positive + false\ positive} \tag{2.9}$$

Presisi mengukur proporsi identifikasi positif yang benar-benar akurat. Presisi penting dalam skenario di mana positif palsu sangat tidak diinginkan(Kaur dan Kaur Sandhu, 2023).

$$recall = \frac{true\ positive}{true\ positive + false\ negative} \tag{2.10}$$

Recall, juga dikenal sebagai sensitivitas, mengukur proporsi kasus positif aktual yang teridentifikasi dengan benar. Hal ini krusial dalam aplikasi di mana hasil negatif palsu merugikan(Kaur dan Kaur Sandhu, 2023).

$$f1 = 2 \times \frac{precission \times recall}{precission + recall} \tag{2.11}$$

F1-skor merupakan rata-rata harmonik presisi dan perolehan, memberikan ukuran kinerja model yang seimbang(Kaur dan Kaur Sandhu, 2023).

BAB III
METODOLOGI PENELITIAN

3.1 Metode Penelitian

Metode eksperimen komputasional berbasis teknologi pemrosesan bahasa alami (*Natural Language Processing/NLP*) dan pembelajaran mesin (*Machine Learning*) digunakan dalam penelitian ini untuk menggunakan pendekatan kuantitatif deskriptif. Untuk tujuan penelitian ini, model bahasa berbasis *transformer IndoBERT* digunakan. Model ini telah dioptimalkan untuk pemahaman dan pemrosesan teks berbahasa Indonesia. Tujuan utama penelitian adalah untuk membuat sistem yang dapat mengidentifikasi *gender* pengguna media sosial dan melakukan klasifikasi sentimen terhadap cuitan dengan tagar #TolakRevisiUUTNI di Twitter.

Diharapkan bahwa metode ini akan memungkinkan untuk mendapatkan gambaran yang lebih sistematis tentang cara publik melihat kebijakan pemerintah yang dibahas dalam tagar tersebut. Selain itu, tujuan penelitian ini adalah untuk melihat perbedaan dalam respons sentimen berdasarkan *gender*. Hal ini dilakukan untuk memberikan pemetaan yang lebih baik tentang bagaimana masyarakat digital melihat masalah kebijakan yang dianggap kontroversial atau menantang.

Untuk mencapai tujuan ini, penelitian ini dirancang dengan menggunakan berbagai model eksperimen. Model 1 menggunakan data murni tanpa perlakuan apa pun, Model 2 menggunakan data yang telah melalui proses *balancing* dengan metode *oversampling*, dan Model 3 hingga Model 5 adalah hasil eksperimen *hybrid* antara *oversampling* dan *undersampling* dengan variasi 10% hingga 30%. Tujuan dari perbedaan desain model ini adalah untuk melihat pengaruh distribusi data terhadap performa model sekaligus dampaknya terhadap fairness dalam klasifikasi.

Hasil evaluasi yang membandingkan performa dan fairness antar model disajikan dalam bentuk tabel dan grafik, selain sistem klasifikasi berbasis IndoBERT. Dengan adanya luaran dalam bentuk visualisasi tabel dan grafik ini, hasil penelitian diharapkan dapat memberikan gambaran yang lebih jelas tentang *trade-off* antara performa dan *fairness* pada setiap model. Ini diukur dengan metrik

standar seperti akurasi, presisi, recall, dan skor F1. Di sisi lain, *fairness* diukur dengan metrik seperti paritas demografi (DP), rasio pilihan (SR), rasio positif asli (TPR), dan rasio positif palsu (FPR).

3.1.1 Studi Literatur dan perumusan masalah

Penulis melakukan studi literatur pada penelitian penelitian terdahulu guna memahami cara kerja, hasil, serta efektivitas penggunaan metode. Dalam penelitian terdahulu berguna dalam memberikan landasan konseptual dalam upaya membangun sistem klasifikasi bagi gender dan sentiment serta mendukung proses evaluasi model scara adil dan akurat dalam judul ini. Kajian kajian tersebut mengenai keterkaitan tentang mengapa gender berpengaruh dalam pengambilan keputusan di kebijakan, tata cara melakukan pelabelan dalam membuat model prediksi gender dan lainnya .

Sedangkan perumusan masalah memiliki Tujuan untuk mendapatkan pemahaman yang lebih baik tentang metode yang paling cocok, serta untuk menciptakan pertanyaan dan batasan masalah penelitian yang lebih spesifik. Tahap ini menghasilkan temuan yang digunakan untuk membangun kerangka kerja penelitian.

3.1.2 Analisis Kebutuhan

Pada tahapan ini melakukan dan mencatat apa saja kebutuhan kebutuhan guna menunjang prosess pengerjaan model dan sistem seperti *software,* Bahasa pemrograman, dan *library* penunjang pengerjaan. Berikut merupakan kebutuhan dalam pembangunan model:

1. Bahasa Pemrograman
   a. *Python*

      Bahasa utama yang digunakan dalam membangun model mulai dari pengolahan data, pelatihan model, dan visualisasi evaluasi hasil. Alasan dipilihnya Bahasa ini adalah kemampuannya dalam penggunaan fitur fitur yang cukup bagus dan di dukung *library open-source* untuk keperluan *machine learning*
2. *Software* (perangkat lunak)
   a. *Jupyter notebook / google collab*

Merupakan sebuah platform pengembangan dengan fitur interakif yang memudahkan penggunanya dalam menulis kode, dokumentasi dan eksplorasi data. *Google collab* juga menyediakan fitur gratis harian penggunaan *GPU* yang membantu mempercepat dalam melakukan prosess pelatihan model berbasis *Transformer* seperti *IndoBERT.*

3. Pusataka dan *framework*
   a. *HuggingFace*

      Merupakan sebuah platform berbasis *cloud* yang berfungsi sebagai tempat penyimpanan hasil pelatihan model untuk dapat dimuat kembali tanpa harus melatih ulang modelnya lagi, selain itu platform ini juga berfungsi sebagai *framework* untuk melakukan *fine-tuning* model *IndoBERT* dengan *indobenchmark/indobert-base-p1*.

   b. *Sklearn(scikit-learn)*

      Digunakan dalam hal *preprocessing*, pelatihan *baseline* model dan perhitungan metriks metriks evaluasi seperti *accuracy, f1-score, recall* dan *precision.*

   c. *Fairlearn*

      Sebuah pustaka untuk melakukan dan mitigasi *bias* berdasarkan atribut sensitif dengan pendekatan *fairness metrics* seperti *Demographic parity* dan *Equalized Odds.*

   d. *Pandas* dan *NumPy*

      Sebuah pustaka yang berfungsi untuk memanipulasi data dan memvisualisasikannya.

   e. *Matplotlib* dan *Seaborn*

      Hamper mirip seperti *pandas* dan *numPy* tetapi pustaka ini khusu untuk melakukan visualisasi seperti visualisasi metriks hasil pelatihan model dan hasil *fairness* metriks.

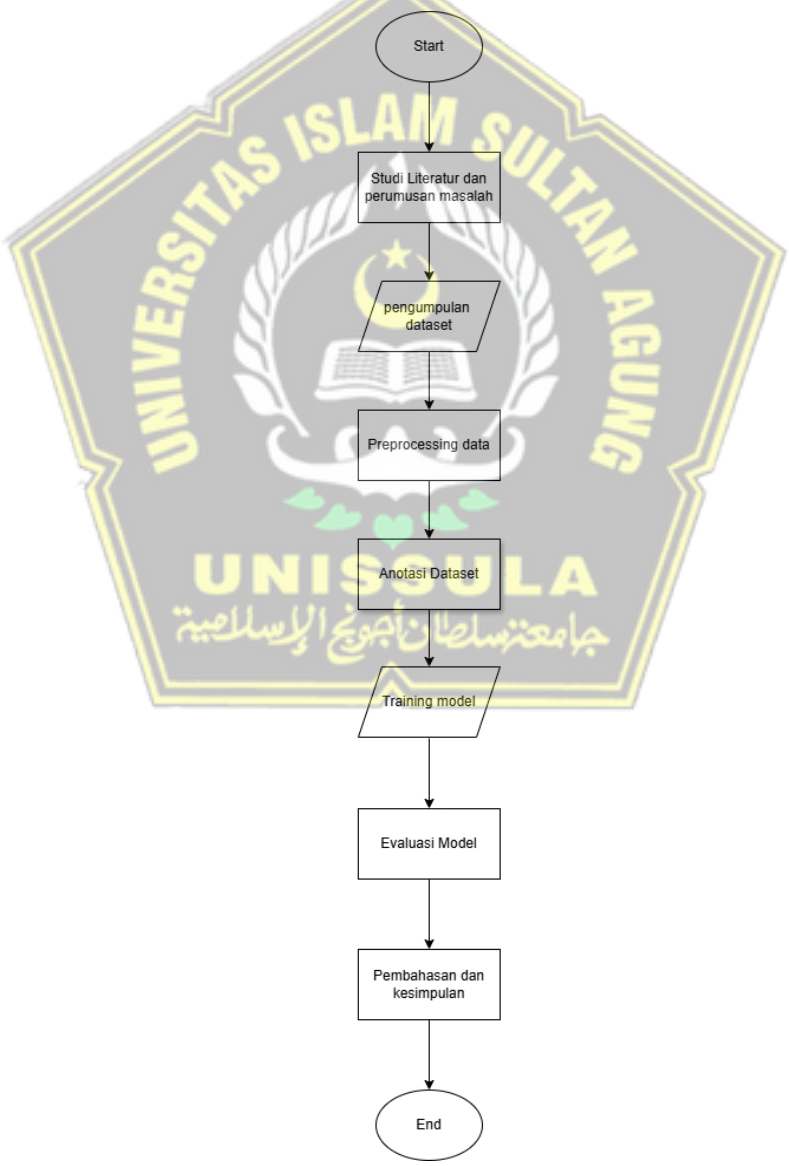
4. Kebutuhan Data
   a. Twitter

      Merupakan sebuah platform media sosial yang banyak digunakan oleh masyarakat luas seluruh dunia, tidak hanya berfungsi sebagai media social

platform ini juga banyak digunakan oleh banyak orang dalam menyampaikan aspirasi, opini maupun kritikan baik kepada perseorangan/individu dan kelompok(bisa organisasi atau pemerintah). Pada tahap ini twitter merupakan sumber pengumpulan data karena menjadi sebuah platfor ajang penyampaian aspirasi dan kritik terhadap pemerintah terutama terhadap isu kontroversial yang di tetapkan oleh pemerintah.

3.1.3 Desain Penelitian

Metode yang akan digunakan dalam melakukan penelitian pada tema ini yaitu *indoBERT*, adapun langkah langkah yang akan di tempuh oleh penulis antara lain:



Gambar 3. 1 *Flowchart* desain penelitian

Penelitian ini bertujuan untuk membangun model klasifikasi berbasis IndoBERT untuk mengevaluasi sentimen dan *gender* dalam hubungan Twitter dengan tagar #TolakRevisiUUTNI. Gambar 3.1 menunjukkan alur penelitian yang digunakan untuk menyusun desain, yang mencakup langkah-langkah penting mulai dari perumusan masalah hingga evaluasi model dan penarikan kesimpulan. Berikut adalah penjelasan untuk setiap tahapan.

Pada tahap awal, penelitian literatur dan perumusan masalah dilakukan dengan tujuan untuk mengumpulkan teori tentang pemrosesan bahasa alami (NLP), model transformasi, terutama *IndoBERT*, metode *balancing* data, dan evaluasi *fairness* dalam klasifikasi teks. Tahap ini juga menjadi dasar untuk merumuskan masalah penelitian, yaitu bagaimana membuat model klasifikasi sentimen dan *gender* yang tidak hanya akurat tetapi juga adil dalam memprediksi kelompok *gender* yang berbeda.

Selanjutnya, pengumpulan dataset dilakukan dengan crawling cuitan Twitter dengan tagar #TolakRevisiUUTNI. Untuk memastikan representasi konteks sosial-politik yang relevan, dataset dikumpulkan dalam jangka waktu tertentu. Data yang dikumpulkan, yang terdiri dari teks cuitan dan metadata akun, kemudian diproses untuk menganotasikan gender dan sentimen.

Preprocessing data adalah tahap berikutnya, yang memastikan model IndoBERT dapat diproses secara optimal. Ini mencakup pembersihan teks dari karakter khusus, normalisasi kata, penghapusan *stopword*, tokenisasi, dan standarisasi format data.

Setelah preprocessing, data melalui dua tahap anotasi anotasi *gender* dan anotasi sentimen. anotasi *gender* dilakukan secara manual menggunakan nama pengguna, *username*, dan deskripsi akun, yang kemudian diverifikasi secara manual. Untuk melabelkan sentimen, teks diannotasi secara manual menggunakan kategori positif, netral, dan negatif.

Selanjutnya, pelatihan model menggunakan IndoBERT dilakukan. Pada titik ini, lima skenario eksperimen berbeda dilakukan. Berikut tabel skenario model yang akan di jalankan:

Tabel 3. 1 Tabel model uji coba

Nama	Keterangan	Pria(%)	Wanita(%)
Model 1	Model dengan pengujian data murni	86.36	13.64
Model 2	Model dengan penanganan penyeimbangan distribusi seluruh data	50	50
Model 3	Model dengan penanganan penggabungan penambahan dan pengurangan sebanyak 10%	65.53	34.47
Model 4	Model dengan penanganan penggabungan penambahan dan pengurangan sebanyak 20%	56.83	43.17
Model 5	Model dengan penanganan penggabungan penambahan dan pengurangan sebanyak 30%	48.57	51.43

Model pertama, yang dikenal sebagai *Baseline* Data Murni menggunakan data asli tanpa *balancing*. Kelima model ini digunakan untuk menguji pengaruh strategi penyeimbangan data terhadap kisaran data. Model 2 (*oversampling*) menyeimbangkan distribusi data dengan teknik *oversampling*. Model 3 (*hybrid resampling* ±10%) menggabungkan *oversampling* dan *undersampling* dengan penyesuaian distribusi ±10%. Model 4 (*hybrid resampling* ±20%) variasi *hybrid* dengan tingkat penyesuaian lebih tinggi, yaitu ±20%. Model 5 (*hybrid resampling* ±30%) penyeimbangan dengan penyesuaian distribusi ±30%.

Selanjutnya, evaluasi model dilakukan dengan dua dimensi utama. Ini terdiri dari metrik performa, yang mencakup skor F1, akurasi, presisi, dan recall. Dimensi kedua adalah metrik kesetaraan, yang mencakup perbedaan kesetaraan demografi (DPD), pilihan rasio (SR), dan metrik kesetaraan lainnya yang relevan untuk mengukur seberapa baik model memberikan prediksi yang seimbang antar gender.

Selanjutnya, hasil evaluasi disajikan dalam bentuk tabel dan grafik yang menunjukkan perbandingan performa dan fairness antar model. Untuk memudahkan analisis perbedaan hasil eksperimen, visualisasi ini diharapkan dapat menunjukkan perbedaan antara akurasi dan fairness yang muncul dari berbagai strategi balancing data yang berbeda.

Dalam langkah terakhir, yang menjawab rumusan masalah penelitian dan menginterpretasikan hasil eksperimen, diskusi dan penarikan kesimpulan dilakukan. Dalam proses ini, penelitian menekankan manfaat penelitian dalam memberikan pemahaman tentang perbedaan distribusi sentimen berdasarkan gender dan pentingnya mempertimbangkan fairness saat membangun model klasifikasi berbasis natural language processing.

3.2 Pengumpulan data penelitian

Peneliti mengumpulkan data pada twee*t* di X(twitter), dimana peneliti melakukan metode scrapping data berdasarkan tema yang diambil. Dimana data yang di dapatkan terdiri dari text, tanggal dibuat, jumlah like, jumlah *retweet* dan masih banyak lagi. Tagar #TolakRevisiUUTNI sebagai kata kunci utama, data penelitian diambil dari Twitter, platform media sosial.dengan rentang waktu pengambilan data dimulai dari awal kemunculan isu revisi tersebut dari 1 februari sampai dengan penelitian ini di buat. Data dikumpulkan dengan *crawling* atau *scraping* menggunakan *API* Twitter atau *tools* pihak ketiga seperti *tweet harvest*. Data ini terdiri dari *username*, tanggal, teks tweet, dan metadata lainnya yang tersedia secara publik. Hanya tweet yang dipilih dalam bahasa Indonesia yang akan dipelajari lebih lanjut.

Pada Bab ini akan membahas mengenai evaluasi dan pembahasan pemaparan hasil yang diperoleh dari penelitian yang berjudul “Analisis *Fairness* model klasifikasi sentiment public terhadap isu revisi UU TNI menggunakan *IndoBERT* berdasarkan *demographic parity* pada kelompok *gender”* yang memiliki keluaran klasifikasi sentiment dengan pendekatan *fairness.* Penelitian ini dilakukan dengan melakukan 5 uji coba model dimana uji coba tersebut diantara lain seperti penggunaan data latih murni, menggunakan data latih dengan balancing, hybrid

balancing 10%-30%.selain itu dilakukan evaluasi hasil pelatihan dengan matriks *Accuracy, Precission, Recall* dan *F1-score* untuk mengukur seberapa baik model dapat melatih model tersebut, selain itu mitigasi *bias* menggunakan pendekatan *fairness* dengan melakukan evaluasi matriks *Accuracy, Selection Rate, True Positve, False positive* lalu dengan pendekatan metode *Demographic parity* untuk mitigasi *Bias* tersebut.

Dataset yang digunakan terdiri dari tweet yang telah melalui prosess pelabelan dalam tiga kelas yaitu *Positive, Negative, Neutral* selain pelabelan sentimen data tweet yang diperoleh tersebut dilabeli dengan informasi *Gender* pengguna tweeter. Tabel 3.2 merupakan distirbusi data latih jumlah tweet berdasarkan label sentimen dan *Gender.* Sedangkan pada Tabel 3.3 merupakan tabel distribusi data test yang akan digunakan.

Tabel 3. 2 Distribusi data Latih

Sentimen	**Laki- Laki**	**Perempuan**	**Total**
Positive	565	254	819
Negative	1167	292	1459
Neutral	185	106	290
Total	1916	652	2568

Tabel 3. 3 Distribusi data test

Sentimen	**Laki- Laki**	**Perempuan**	**Total**
Positive	525	87	612
Negative	670	101	771
Neutral	58	2	60
Total	1253	190	1443

Pada tabel 3.2 dan 3.3 merupakan distribusi data latih dan data test yang berhasil di peroleh dari tweet-tweet pada twitter. Dari data data yang diperoleh kemudian dilakukan anotasi oleh 2 anotator laki-laki dan perempuan untuk melabeli pada sentimen dan juga *gender,* anotasi ini bersifat netral jadi merepresentasikan apa yang dinilai oleh annotator, kemudian hasil dari anotasi tersebut di samakan dan diambil hasil yang memiliki hasil yang sama dari kelas sentimen dan *gender*.

Tabel 3. 4 pengumpulan data

created_at	full_text	username	favorite_count	gender	sentimen
Sat Apr 19 04:35:14 +0000 2025	Turut berduka untuk korban semoga amal ibadahnya diterima di sisi Nya. amiin minta tolong bantu ramein teman kami mati dibunuh TNI!! @BudiBukanIntel @aromapetrikorr @barengwarga @TxtSerang #TolakRevisiUUTNI #TolakRUUPolri	yasinsyahid1	16131	Laki-laki	negatif
Thu Apr 17 14:17:13 +0000 2025	Indonesia Gelap! #GagalkanUUTNI #CabutUUTNI #TolakUUTNI #TolakRevisiUUTNI #PeringatanDarurat #IndonesiaGelap #TolakDwifungsiABRI #TolakRUUPolri #TolakRUUKejaksaan #SupermasiSipil #PerempuanMelawan	01iann	0	Laki-laki	Negatif

created_at	full_text	username	favorite_count	gender	sentimen
Thu Apr 17 08:35:48 +0000 2025	Hallo teman teman. Aku mau sharing info kalau artikel tentang #TolakRUUPolri dan #TolakRevisiUUTNI yang ditulis oleh @KasperNollet sudah dipublikasi melalui media @mondiaalnieuws. Jangan lupa dibaca dan share sebanyak mungkin. Cc. @barengwarga @BudiBukanIntel @karimakayyim	SeriDianawati	161	perempuan	positif

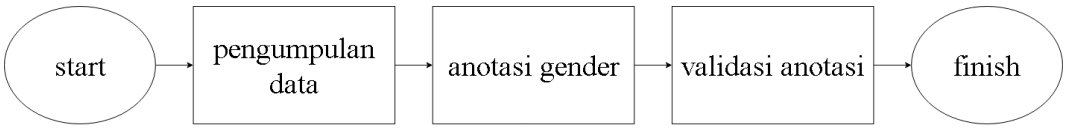
3.3 Anotasi Data

Pada tahapan ini melakukan anotasi pada setiap dataset agar dapat di klasifikasikan berdasarkan tiap dataset. Setiap entri dalam dataset diberi label melalui anotasi data, yang digunakan sebagai *ground truth* selama proses pelatihan dan evaluasi model. Anotasi difokuskan pada dua aspek anotasi sentimen dan anotasi *gender*. Keduanya penting karena tujuan penelitian ini adalah membangun model klasifikasi sentimen serta menganalisis perbedaan distribusi sentimen berdasarkan gender.

3.3.1 Anotasi Gender

Gender pengguna Twitter yang mencuit dengan tagar #TolakRevisiUUTNI diidentifikasi melalui anotasi *gender*. Dua anotator dengan kategori label laki-laki dan perempuan melakukan pelabelan secara manual. Beberapa sumber informasi dapat digunakan untuk menentukan *gender*, seperti nama akun dan username yang menunjukkan gender tertentu deskripsi profil pengguna yang mengandung kata ganti atau keterangan diri dan konten cuitan dalam kasus di mana ada indikasi

linguistik yang menunjukkan identitas gender. Dalam kasus yang ambigu, kedua anotator berbicara untuk mencapai kesepakatan. Untuk menjaga kualitas anotasi, data yang tidak dapat ditentukan *gender* nya dikeluarkan dari dataset.

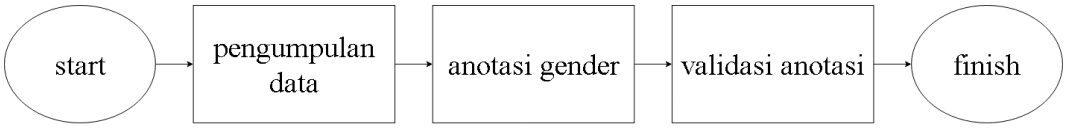


Gambar 3. 2 *Flowchart* anotasi gender

Setiap teks ditinjau menggunakan informasi penulis, seperti metadata akun atau deskripsi profil, dan kemudian dilabelkan sesuai kategori yang telah ditetapkan. Gender biasanya terbagi menjadi dua kategori laki-laki (0) dan perempuan (1). Untuk menetapkan label ini, annotator dapat melakukannya secara manual atau dengan menggunakan aturan tertentu, seperti identifikasi nama atau ciri linguistik. Untuk menjamin keakuratan, hasil anotasi divalidasi melalui cross-check antar-annotator. Sebagai hasil dari proses ini, dataset yang memiliki label gender yang konsisten telah diselesaikan dan siap untuk digunakan untuk analisis lebih lanjut.

3.3.2 Anotasi Sentiment

Selain itu, anotasi sentimen dilakukan untuk membagi isi cuitan ke dalam tiga kategori yaitu positif, negatif, dan netral. Pelabelan dilakukan secara manual oleh dua anotator yang dipilih dalam kondisi netral untuk menghindari *bias* terhadap salah satu kategori sentimen. Menurut definisi operasional, label ditetapkan sebagai berikut cuitan dikategorikan positif jika mengandung dukungan, persetujuan, atau emosi positif terhadap masalah cuitan dikategorikan negatif jika mengandung penolakan, kritik, atau emosi negatif dan cuitan dikategorikan netral jika tidak secara eksplisit menyatakan dukungan atau penolakan, tetapi tetap informatif atau netral. Hasil anotasi divalidasi melalui proses perbandingan hasil antar-anotator untuk meningkatkan kualitas. Jika ada perbedaan pendapat, diskusi dilakukan untuk mencapai konsensus. Hasil anotasi digunakan sebagai data berlabel untuk pelatihan model *IndoBERT*.



Gambar 3. 3 Flowchart anotasi sentimen

Setelah anotasi gender selesai, anotasi sentimen terhadap teks dilakukan. Ekspresi emosional atau pendapat yang terkandung dalam setiap kalimat atau tweet diperiksa. Membagi data ke dalam tiga kategori utama dilakukan untuk anotasi: negatif (-1), netral (0), dan positif(+1). Annotator secara manual menilai konten teks dengan mempertimbangkan makna kata, konteks kalimat, dan nuansa emosional. Uji kesepakatan antar-annotator juga dilakukan untuk meningkatkan reliabilitas dan memastikan bahwa hasil anotasi dapat dianggap konsisten. Proses ini menghasilkan dataset dengan label sentimen yang jelas. Data ini akan digunakan sebagai dasar untuk pelatihan dan evaluasi model klasifikasi berbasis sentimen di masa mendatang.

3.4 Preprocessing data

Preprocessing, tahap penting dalam pemrosesan bahasa alami (NLP), bertujuan untuk membersihkan dan menormalkan teks agar model dapat memprosesnya dengan benar. Dalam penelitian ini, prosedur *preprocessing* termasuk:

Gambar 3. 4 Flowchart Preprocessing

1. *Casefolding*

   Pada tahap ini, seluruh huruf dalam teks tweet diubah menjadi huruf kecil (*lowercase*). Hal ini dilakukan untuk menghindari redundansi kata akibat perbedaan kapitalisasi.dan memudahkan model untuk belajar dengan teks tersebut.

2. Pembersihan teks

   Pada tahap ini dilakukan penghapusan kata kata yang tidak di perlukan, seperti *URL*, mention, tanda baca, angka dan karakter karakter khusus yang tidak di perlukan.

3. Normalisasi kata

   Normalisasi adalah proses mengubah kata-kata yang tidak baku, kata-kata alay, atau singkatan menjadi bentuk yang tepat. Ini dapat dilakukan dengan menggunakan alat pengolahan bahasa natural (NLP) Indonesia seperti Sastrawi, KamusKataAlay, atau kamus manual. Kamus kamus kata tersebut dapat di temukan di internet.

4. *Tokenisasi*

   Tahapan ini bertujuan untuk memiasahkan dari sebuah kalimat menjadi potongan kata kata. Hal ini bertujuan Menyiapkan data agar lebih bersih dan siap digunakan dalam pemodelan. Dalam kasus ini penulis menggunakan metode tokenisasi dari *indoBERT*, yang membedakan tokenisasi ini dengan tokenisasi lainnya, yang memecah teks menjadi bagian-bagian kecil yang disebut token, yang biasanya terdiri dari kata. Sedangkan tokenisasi dengan *indoBERT* karena Tokenisasi *WordPiece IndoBERT* berbeda dari tokenisasi konvensional karena berbasis sub-kata yang bertujuan untuk menangani kata yang tidak dikenal

5. *Stopword Removal*

   Menghapus kata-kata umum yang tidak berpengaruh terhadap analisis (menggunakan daftar *stopword* bahasa Indonesia). Kata-kata umum yang sering muncul dalam teks tetapi tidak signifikan saat dinilai, seperti "yang", "dan", "di", "ke", dan "dengan," dihapus agar hanya kata-kata bermakna atau berbobot emosional atau informatif yang tersisa.

6. *Stemming*

   salah satu tahapan penting dalam preprocessing teks yang bertujuan untuk mengubah kata berimbuhan menjadi bentuk dasarnya (root word atau kata dasar) dengan menghapus awalan, akhiran, sisipan, atau kombinasi imbuhan lainnya. Misalnya, istilah-istilah seperti "berlari", "lari-lari", dan "pelari" semuanya direduksi menjadi istilah dasar "lari".

3.5 *Training* Model

Pada tahap pelatihan Model metode yang digunakan adalah *IndoBERT*, yang dikenal sebagai *indobenchmark/indobert-base-p1*, merupakan sebuah jenis *BERT*

yang dilatih khusus untuk bahasa Indonesia.Pelatihan dijalankan dengan Dataset dibagi menjadi data latihan dan data uji dengan rasio 2658 data latih dan 1443 data test. Selama proses eksplorasi, parameter pelatihan seperti tingkat pembelajaran, *learning rate* 2e-5, ukuran 12 *batch*, dan jumlah 3*epoch*. Setelah melakukan pelatihan pada model, model dapat di simpan di *HuggingFace* tempat penyimpanan berbasis *cloud* yang terkhususkan untuk model model kecerdasan buatan agar bisa di gunkana kembali nantinya. Dalam proses diatas menggunakan metode *indoBERT* dimana metode tersebut memiliki tahapan-tahapannya diantara lain:

1. *Input Text* (Tweet)

   Berisi teks tweet mentah dan setiap tweet adalah representasi dari opini yang akan dijadikan dasar prediksi gender dan juga sentiment.

2. *IndoBERT Encoding* (*Transformer Layers*)

   Pada *indoBERT* memilik 12 layers *Transformers* dimana pada tiap layernya memproses representasi dari token dengan *self-attention* untuk menghasilkan sebuah pemahaman konteks antar kalimat.

3. *Classification Head*

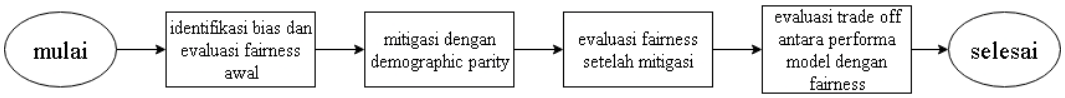
   Output dari token akan dimasukkan ke layer linear biasa (*fully connected layer*), yang bertujuan mengubah representasi kalimat menjadi skor untuk masing masing label.

4. *Evaluation*

   Diakhir *epoch/train* model akan di uji data validasi yang hasilnya akan di analisis dengan metric seperti *Accuracy, f1-Score, Precission, Recall.*

3.6 Mitigasi *bias* dan penerapan *Fairnsess*

Bias dalam penelitian pembelajaran mesin adalah masalah penting yang harus diperhatikan karena dapat menyebabkan hasil prediksi yang tidak adil. Contoh Misalnya, jumlah data dari kelompok gender tertentu lebih dominan daripada kelompok gender lain. Karena ketidakseimbangan tersebut, model dapat mempelajari pola dari kelompok mayoritas dengan lebih baik. Akibatnya, prediksi yang dibuat untuk kelompok minoritas menjadi kurang akurat.



Gambar 3. 5 flowchart mitigasi dan penerapan fairness

Pada awal proses mitigasi bias dan penerapan keadilan, penelitian ini dimulai dengan identifikasi bias dan evaluasi keadilan awal. Pada tahap ini, prediksi model dievaluasi untuk mengetahui seberapa mungkin model menunjukkan bias terhadap atribut sensitif, khususnya gender. Untuk menunjukkan adanya perbedaan per demografi, evaluasi keadilan awal dilakukan dengan menggunakan metrik seperti tingkat pemilihan, perbedaan paritas demografi, dan rasio paritas demografi. Setelah bias ditemukan, langkah berikutnya adalah menguranginya dengan menggunakan metode paritas demografis. Metode post-processing digunakan untuk melakukan mitigasi ini. Tujuan algoritma Fairlearn ThresholdOptimizer adalah untuk mengubah ambang batas keputusan model agar hasil prediksi lebih seimbang antar kelompok sensitif. Setelah mitigasi, tahap selanjutnya adalah evaluasi fairness setelah mitigasi. Pada tahap ini, metrik fairness dihitung kembali untuk mengevaluasi seberapa efektif mitigasi dalam mengurangi disparitas prediksi antara kelompok gender. Hasil evaluasi ini memberikan gambaran tentang seberapa jauh bias dapat dikurangi setelah implementasi teknik mitigasi. Terakhir, evaluasi hasil antara efisiensi model dan keadilan dilakukan. Pada tahap ini, untuk mengevaluasi dampak dari mitigasi, metrik kinerja model (tepat, presisi, recall, dan skor F1) dibandingkan sebelum dan sesudah mitigasi. Oleh karena itu, agar solusi yang dihasilkan tetap relevan, analisis tidak hanya melihat peningkatan fairness, tetapi juga mengevaluasi bagaimana performa model dan fairness seimbang.

Penelitian ini menggunakan strategi mitigasi yang didasarkan pada prinsip Demographic Parity (DP) untuk mengatasi bias. Konsep ini menekankan bahwa kemungkinan mendapatkan prediksi tertentu, seperti sentimen positif, tidak boleh dipengaruhi oleh atribut sensitif seperti gender. Dengan kata lain, model dikatakan adil apabila proporsi prediksi antara laki-laki dan perempuan seimbang. Mitigasi dilakukan dengan menyeimbangkan distribusi data dengan menggunakan metode balancing seperti oversampling, undersampling, atau hybrid resampling. Tujuan dari pendekatan ini adalah agar model dapat belajar dari data yang lebih

representatif, sehingga mengurangi kemungkinan keberpihakan terhadap kelompok tertentu.

Dalam penelitian ini, tiga metrik utama digunakan untuk menilai fairness: Selection Rate (SR), Demographic Parity Difference (DPD), dan Demographic Parity Ratio (DPR). SR mengukur proporsi prediksi positif yang diterima oleh masing-masing kelompok, dan DPD menilai selisih nilai SR antar kelompok, dengan nilai yang mendekati nol menunjukkan fairness yang baik. Untuk mengevaluasi kemampuan model untuk menghasilkan prediksi yang setara antar gender sebelum dan sesudah mitigasi, ketiga metrik ini digunakan bersama-sama.

Hasil dari penerapan mitigasi menunjukkan bahwa penanganan data yang timpang dan penerapan strategi mitigasi fairness sangat penting untuk membangun model yang handal. Hasil menunjukkan bahwa distribusi data yang lebih seimbang mampu menurunkan nilai DPD dan mendekatkan nilai DPR ke angka 1, sehingga tingkat keselarasan model meningkat tanpa mengorbankan performa secara signifikan. Memastikan bahwa model memperlakukan setiap kelompok gender dengan adil. Upaya ini mendukung pengembangan sistem analisis berbasis AI yang lebih moral, terbuka, dan dapat dipercaya.

BAB IV
HASIL DAN ANALISIS PENELITIAN

4.1 *Preprocessing* Data

Preprocessing, yang dilakukan sebelum proses pelatihan model, dilakukan untuk meningkatkan kualitas data teks agar algoritma pembelajaran mesin lebih mudah memahaminya. Preprocessing adalah proses yang dilakukan melalui berbagai langkah. Pertama, teks dibersihkan dengan menghapus kata imbuhan, URL, emoji, dan karakter khusus yang tidak relevan dengan analisis. Agar model tidak membedakan kata yang sama dengan format penulisan yang berbeda, seluruh teks diubah menjadi huruf kecil. Normalisasi juga digunakan untuk menyeragamkan bentuk kata yang memiliki arti yang sama tetapi ditulis dengan cara yang berbeda, seperti kata tidak baku atau singkatan. Selain itu, tokenisasi digunakan untuk memecah kalimat menjadi bagian-bagian kata yang lebih kecil, yang memudahkan analisis. Tahap preprocessing ini menghasilkan data teks yang bersih, seragam, dan terstruktur yang siap digunakan pada tahap pelatihan model.

Tabel berikut menunjukkan potongan data sebelum dan sesudah preprocessing, yang menunjukkan hasil dari tahapan preprocessing. Tabel ini menunjukkan bagaimana teks mentah yang masih mengandung kata-kata atau simbol yang tidak baku diubah menjadi bentuk yang lebih konsisten dan bersih.

Tabel 4. 1 data preprocessing

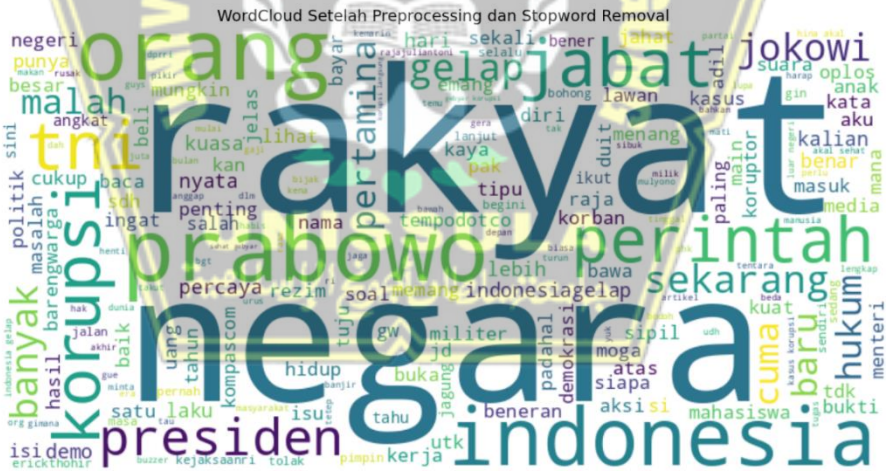
no	sebelum	sesudah
1	Turut berduka untuk korban semoga amal ibadahnya diterima di sisi Nya. amiin minta tolong bantu ramein teman kami mati dibunuh TNI!! @BudiBukanIntel @aromapetrikorr @barengwarga @TxtSerang #TolakRevisiUUTNI #TolakRUUPolri	turut duka korban moga amal ibadah terima sisi amiin minta bantu ramein teman mati bunuh tni budibukanintel aromapetrikorr barengwarga txtserang

no	sebelum	sesudah
2	Hallo teman teman. Aku mau sharing info kalau artikel tentang #TolakRUUPolri dan #TolakRevisiUUTNI yang ditulis oleh @KasperNollet sudah dipublikasi melalui media @mondiaalnieuws. Jangan lupa dibaca dan share sebanyak mungkin. Cc. @barengwarga @BudiBukanIntel @karimakayyim	hallo teman teman aku sharing info artikel tulis kaspernollet publikasi lalu media mondiaalnieuws lupa baca share banyak mungkin cc barengwarga budibukanintel karimakayyim
3	Ingat! Kita ini perintis bukan komunis! Jadi ga perlu ditakuti #TolakRevisiUUTNI #TolakRUUPolri	ingat rintis komunis perlu takut
4	Yg demo dibilang bikin macet terus dibubarin sampe diculik. Kok yg ini ga? #TolakRevisiUUTNI #TolakDwiFungsiAbri	demo macet dibubarin culik
5	temen stay safe buat akun sosmed kalian untuk jaga nyalain verivikasi 2 langkah. gak twitter gak Ig semua akun aku udah gak aman kayaknya. #BatalkanRUUTNI #BatalkanRevisiUUPolri #GagalkanRUUTNI #CabutRUUTNI #TolakRUUTNI #TolakRevisiUUTNI #SupermasiSipil	temen stay safe akun sosmed kalian jaga nyalain verivikasi langkah twitter ig akun aku aman
	banyak pqrt yaa kaa nomi. jangan takut masih ada Tuhan yang adil sama masyarakat yang membela kebenaran dan masyarakat yang waras (kecuali	banyak pqrt yaa kaa nomi takut tuhan adil masyarakat bela benar masyarakat waras buzzer duit doang fuck renta tegak adil

no	sebelum	sesudah
	buzzer yang mau duit doang) fuck pemerenta tegakkan keadilan #TolakRevisiUUTNI #TolakRUUTNI https://t.co/vJ491d8Wvx	

Tabel 4.1 menunjukkan bahwa teks yang sebelumnya mengandung singkatan, tanda baca, dan emotikon telah diproses secara efektif menjadi bentuk yang lebih sederhana yang tetap sesuai dengan kaidah bahasa formal. Hal ini memastikan bahwa kata-kata yang dipelajari selama proses pemodelan adalah kata-kata yang benar-benar menggambarkan isi pesan, bukan sekadar variasi dalam penulisan.

Setelah tahap preprocessing, visualisasi juga diperlukan untuk menunjukkan distribusi kata yang paling sering muncul dalam dataset. Wordcloud, representasi visual yang menampilkan kata-kata dominan dalam bentuk awan kata, adalah salah satu visualisasi yang paling umum digunakan. Jumlah kali sebuah kata muncul dalam data berkorelasi positif dengan ukuran wordcloudnya.



Gambar 4. 1 Visualisasi akhir dengan *worldcloud*

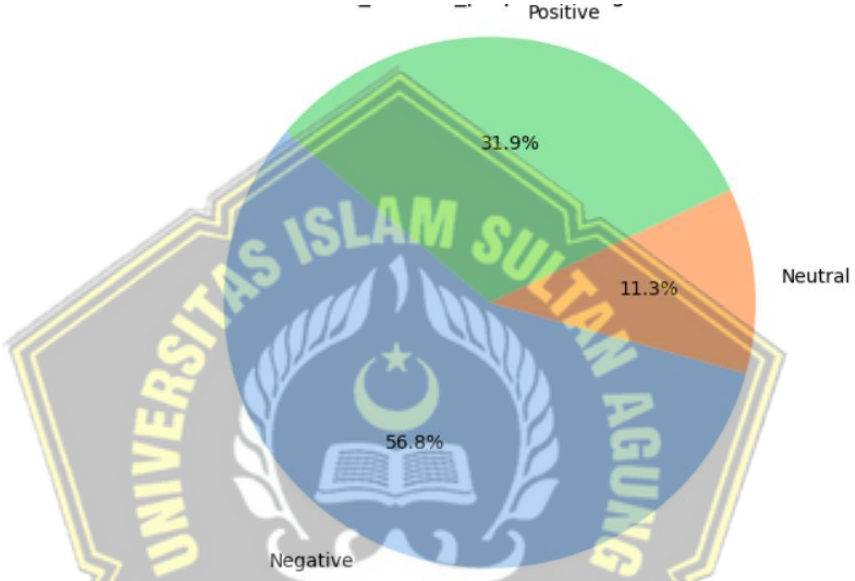
Gambar 4.1 wordcloud menunjukkan bahwa kata-kata yang sering muncul, seperti "rakyat", "negara", "korupsi", dan "presiden," terkait erat dengan subjek penelitian. Ini menunjukkan bahwa proses preprocessing berhasil menyimpan kata-kata kunci penting untuk analisis sentimen. Oleh karena itu, wordcloud berfungsi sebagai representasi visual yang membantu menjelaskan isi data dan menjadi bukti bahwa data sudah siap untuk digunakan saat tahap pemodelan selesai.

4.2 Hasil pemodelan dan evaluasi

4.2.1 Model 1 (Model Murni Tanpa Adanya Penyeimbangan Data)

1. Visualisasi distribusi data

Pada tahap ini percobaan pertama menggunakan dataset latih yang masih murni tanpa adanya prosess balancing data ataupun *reweighting.* Dimana distribusi dataset sebagai berikut:



Gambar 4. 2 Distribusi data latih model 1

Pada gambar 4.2 menampilkan visualisasi mengenai distribusi data latih pada model satu dimana dapat dilihat bahwa dominasi data terbanyak pada data negative selanjutnya data postif dan netral. sebanyak 56,8% pada data negative, 31,9% pada data positif dan yang terakhir 11.3% data netral.

2. *Finetuning* Model

Tahap selanjutnya setelah memvisualisasikan data merupakan tahapan yang paling penting yaitu *finetung* model tahapan ini merupakan tahapan dimana model belajar dari data latih yang sudah di siapkan tadi untuk kemudian digunakan dalam prediksi pada data *test.*penguunaan Model dasar *indobenchmark/indobert-base-p1*, yang telah dilatih pada korpus bahasa Indonesia, digunakan. *Tokenizer* yang tersedia untuk model digunakan untuk tokenisasi, dan metode *truncation* digunakan untuk memastikan panjang input

sesuai dengan batas model.pelatihan model dimuat dengan *trainer API* dari *Huggingface* menggunakan konfigurasi *TrainingArguments* dengan menggunakan *epoch* 3 dalam pelatihannya, *batch size 16* dan *learning rate* 2e-5.

Alasan menggunakan 3 *epoch* dalam pelatihannya karena jumlah dataset yang terbatas menjadikan 3 *epoch* sudah cukup untuk melaukan pelatihan pada model selain itu dikhawatirkan juga terjadinya *overfitting* dalam pelatihan model tersebut. Untuk menjamin penilaian yang adil pada ketiga kelas sentimen, proses evaluasi menggunakan metrik penting seperti ketepatan, ketepatan, recall, dan skor F1 dengan pendekatan rata-rata makro (*macro averaging*).

3. Evaluasi hasil pelatihan

Setelah dilakukanya prosess pelatihan sebelumnya menggunakan skema *finetuning* terhadap arsitektur *IndoBERT*, menghasilkan metriks evaluasi seperti *accuracy, precision, recall* dan *f1-score.* metriks metriks tersebut berfungsi untuk melihat menilai seberapa baik model dalam melakukan klasifikasi sentimen secara keseluruhan. Berikut penjelasan mengenai metriks metriks tersebut

Dari hasil pelatihan model pertama menghasilkan metriks metriks dengan hasil sebagai berikut:

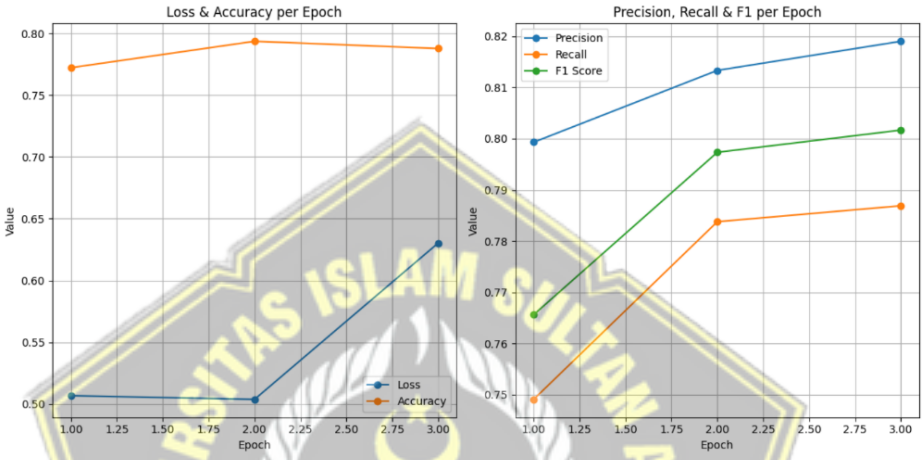
Tabel 4. 2Hasil metriks pelatihan model 1

Validation_loss	***Accuracy***	***Precission***	***Recall***	***F1-Score***
0.5066	0.793774	0.813322	0.783786	0.797335

Tabel 4.2 menyajikan hasil evaluasi model klasifikasi sentimen menggunakan arsitektur IndoBERT. Metrik ketepatan, ketepatan, recall, dan skor F1 digunakan untuk menilai data validasi. Didasarkan pada nilai kehilangan validasi terendah, yaitu 0,5066, model yang paling efektif dapat menghasilkan akurasi sebesar 79,4%, ketepatan sebesar 81,3%, recall sebesar 78,4%, dan skor F1-sebesar 79,7%.

Menurut nilai evaluasi, model telah mencapai titik konvergensi ideal selama proses pelatihan. Sebuah tingkat akurasi yang relatif tinggi dengan

keseimbangan antara ketepatan dan recall menunjukkan bahwa model tidak hanya mampu mengklasifikasikan dengan benar, tetapi juga mampu menangani distribusi kelas yang ada dengan konsisten. Oleh karena itu, memilih model dengan kehilangan validasi terendah dapat dianggap sebagai contoh untuk digunakan pada tahap pengujian berikutnya.

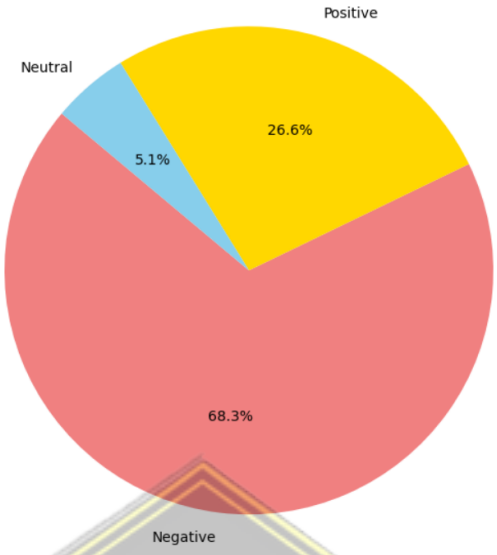


Gambar 4. 3 Grafik visualisasi metriks

Gambar 4.3 merupakan sebuah visualisasi dari hasil pelatihan yang dilakukan di model pertama.

4. Prediksi menggunakan model

Setelah melakukan pelatihan pada data latih dan melakukan evaluasi terhadap hasil pembelajaran dan menyimpannya pada platform *huggingface* kemudian kita memuat ulang dari model yang sudah di simpan untuk digunakan kembali pada prediksi data test yang sudah di siapkan sebelumnya. Dengan menghasilkan prediksi sebanyak 68.29 data negatif, 26.63 data positif dan 5.08 data netral dengan visualisasi sebagai berikut:



Gambar 4. 4 Visualisasi distribusi hasil prediksi

Pada gambar 4.4 merupakan visualisasi distribusi yang dihasilkan dari prediksi dengan model yang sudah dilatih sebelumnya. Model ini merupakan model dengan data latih murni tanpa ada penanganan tambahan seperti *balancing* atau lainnya.

5. Evaluasi *Fairness* dasar dan mitigasi lanjutan *Fairness* dengan *Demographic parity*

Fokus utama penelitian ini, proses evaluasi *fairness*, dibahas secara khusus di tahap ini. Tujuan dari proses evaluasi ini adalah untuk mengevaluasi sejauh mana model klasifikasi sentimen yang dibangun dapat bersikap adil terhadap berbagai kelompok pengguna, dalam hal ini berdasarkan karakteristik *gender*. Pada titik ini, model awal dievaluasi menggunakan metrik *fairness* dasar seperti *accuracy*, *Rate of Selection* (SR), *True Positive Rate* (TPR), dan *False Positive Rate* (FPR).

Tujuan evaluasi ini adalah untuk mendapatkan gambaran awal tentang distribusi performa model antar kelompok. Kemudian, dengan menggunakan pendekatan *fairness* yang lebih sistematis. Hasil evaluasi didasarkan pada output klasifikasi sentimen, yang akan dianalisis lebih lanjut untuk menemukan kemungkinan bias atau ketidakseimbangan prediksi antar *gender*. Hasil evaluasi *fairness* dasar terhadap model awal berikut:

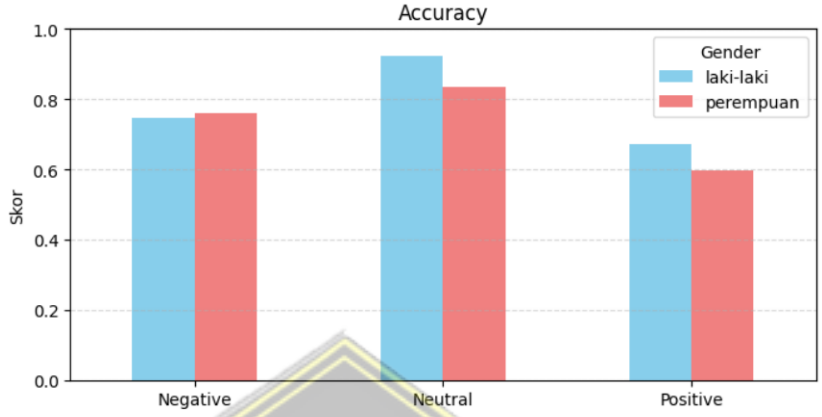
Tabel 4. 3 Metriks fairness dasar

sentimen	Fairness metriks					
	Laki Laki			Perempuan		
	Accuracy	True positive	False positive	Accuracy	True positive	False positive
Positif	0.672	0.452	0.166	0.596	0.436	0.297
Negatif	0.747	0.871	0.403	0.761	0.910	0.367
Netral	0.922	0.081	0.051	0.834	0.0	0.031

Dalam sentimen netral, model tidak dapat menemukan data untuk kelompok perempuan ini ditunjukkan oleh True Positive Rate (TPR) sebesar 0.0, sedangkan TPR untuk kelompok laki-laki adalah 0.081. Selain itu, pada sentimen Positif, False Positive Rate (FPR) untuk Laki-laki lebih tinggi (0,166) dibandingkan Perempuan (0,297), menunjukkan adanya Bias, di mana model tidak memberikan peluang yang sama bagi kelompok perempuan untuk mendapatkan hasil yang benar. Secara keseluruhan, walaupun data Laki-laki lebih sering diklasifikasikan sebagai Positif daripada data Perempuan, ini menunjukkan adanya Bias equal opportunity dan juga predictive bias.

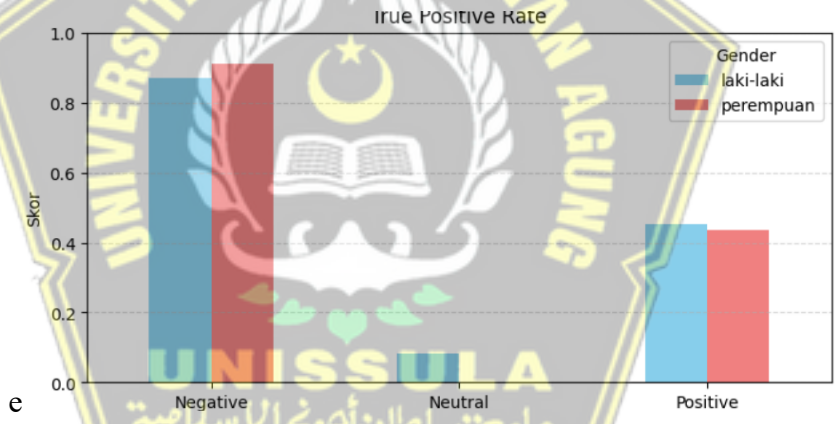
Bias terjadi karena distribusi data yang tidak seimbang, terutama pada kelompok netral perempuan. Akibatnya, model tidak dapat mengidentifikasi pola yang representatif. Selain itu, gaya bahasa yang berbeda yang digunakan orang dari kedua gender memudahkan model untuk mengidentifikasi pola sentimen pada laki-laki. Akibatnya, TPR perempuan sangat rendah, sementara FPR laki-laki pada kelas positif lebih tinggi. Kondisi ini menunjukkan bahwa

model menguntungkan laki-laki dan tidak adil pada perempuan. Akibatnya, terjadi bias kesempatan yang sama dan bias prediktor.



Gambar 4. 5 Visualisasi metriks *Accuracy*

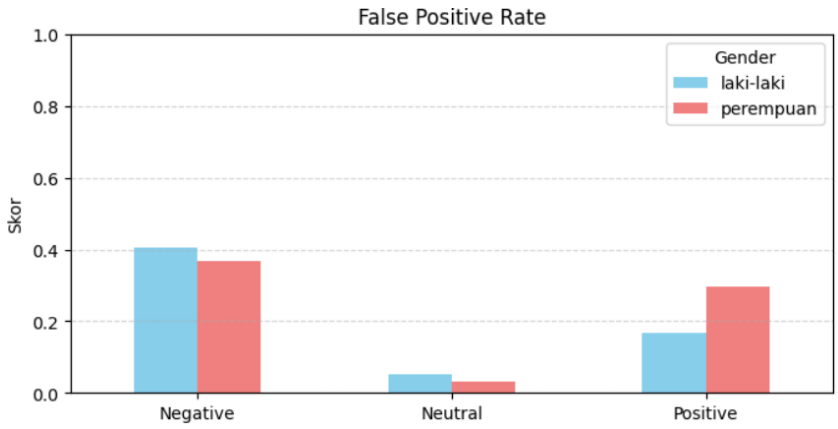
Meskipun tingkat akurasi masih sangat tinggi,masih terdapat bias antar kelompok memengaruhi kinerja model.



e

Gambar 4. 6 Visualisasi metriks TPR

Terdapat perbedaan besar antara kelompok, menurut visualisasi TPR. Ini menunjukkan bahwa kemampuan model untuk membuat prediksi negatif tidak seragam.



Gambar 4. 7 Visualisasi metriks FPR

Grafik FPR menunjukkan perbedaan antara kelompok, dengan kelompok tertentu cenderung lebih sering menerima prediksi yang salah.

Tabel 4. 4 Tabel sebelum mitigasi Demographic Parity

sentimen	Demographic parity					
	Laki Laki			Perempuan		
	Selection rate	Demographic parity difference	Demographic parity ratio	Selection rate	Demographic parity difference	Demographic parity ratio
Negatif	0.660	0.041	1.060	0.619	0.041	1.060
Netral	0.052	0.021	1.686	0.027	0.021	1.686
Positif	0.286	0.019	1.076	0.353	0.019	1.076

Setiap kelas sentimen memiliki tingkat bias yang berbeda, menurut hasil evaluasi fairness dengan metrik paritas demografi. Dalam kelas negatif, nilai pilihan rata-rata (SR) kelompok laki-laki sedikit lebih tinggi daripada perempuan (0,619), dengan perbedaan demografi (DPD) 0,041 dan perbandingan demografi (DPR) 1,060. Nilai-nilai ini menunjukkan bahwa model relatif seimbang dalam memberikan prediksi negatif bagi kedua kelompok gender, sehingga potensi bias rendah. Di kelas netral, sebaliknya, ditemukan indikasi bias yang lebih siginifikan. Dengan DPR sebesar 1,686, SR laki-laki (0,052) hampir dua kali lipat lebih tinggi daripada SR perempuan (0,027), menunjukkan ketidakseimbangan prediksi antar gender.

Kondisi ini menunjukkan bahwa model lebih sering mengklasifikasikan data laki-laki sebagai sentimen netral daripada data perempuan. Di sisi lain, SR

kelompok perempuan (0,353) sedikit lebih tinggi daripada SR kelompok laki-laki (0,286), dan DPR sebesar 1,076 masih dekat dengan nilai ideal (1). Ini menunjukkan bahwa, meskipun ada perbedaan kecil, distribusi prediksi positif antar gender masih relatif seimbang. Oleh karena itu, dapat disimpulkan bahwa, sebelum mitigasi, kelas sentimen netral memiliki bias paling dominan, sementara kelas sentimen negatif dan positif menunjukkan distribusi gender yang lebih adil.

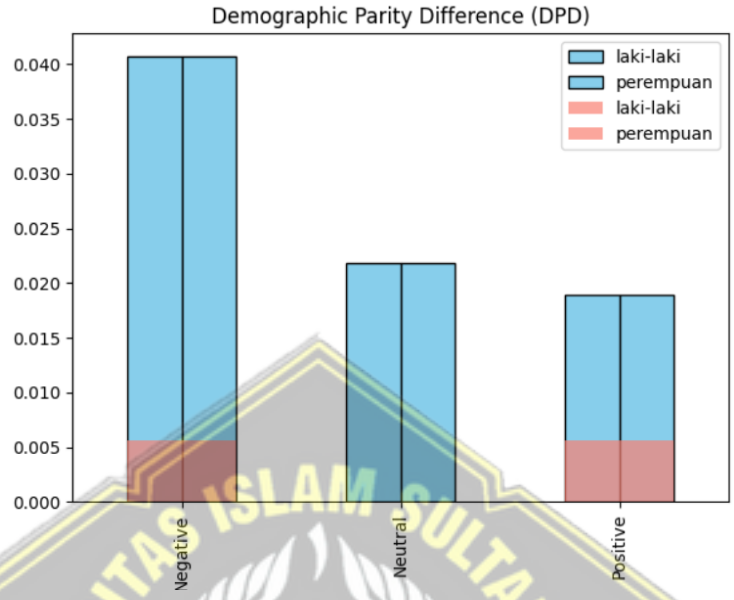
Tabel 4. 5 Setelah mitigasi

sentimen	Demographic parity					
	Laki Laki			Perempuan		
	Selection rate	Demographic parity difference	Demographic parity ratio	Selection rate	Demographic parity difference	Demographic parity ratio
Negatif	0.576	0.005	1.007	0.579	0.005	1.008
Netral	0.0	0.0	1.0	0.0	0.0	1.0
Positif	0.377	0.005	1.011	0.382	0.005	1.014

Berdasarkan hasil evaluasi fairness pasca-mitigasi, distribusi prediksi antar gender telah meningkat secara signifikan. Nilai pilihan rasio (SR) untuk laki-laki (0,576) dan perempuan (0,579) menunjukkan proporsi yang ideal di kelas negatif. Ini diperkuat oleh fakta bahwa bias kelas negatif telah berhasil dikurangi, karena Demographic Parity Difference (DPD) hanya 0,005 dan Demographic Parity Ratio (DPR) sebesar 1,008, keduanya mendekati nilai ideal 1. Selain itu, nilai SR laki-laki (0,0) dan perempuan (0,0) Namun, metrik keadilan menunjukkan peningkatan yang signifikan, dengan DPD 0,0 dan DPR 1,0. Nilai ini menunjukkan bahwa distribusi prediksi netral gender sudah ideal, tanpa indikasi bias. Pola yang sama juga ditemukan di kelas positif SR laki-laki (0,377) dan SR perempuan (0,382) sebanding, dengan nilai DPD 0,005 dan DPR 1,014, yang juga berada di dekat keseimbangan sempurna.

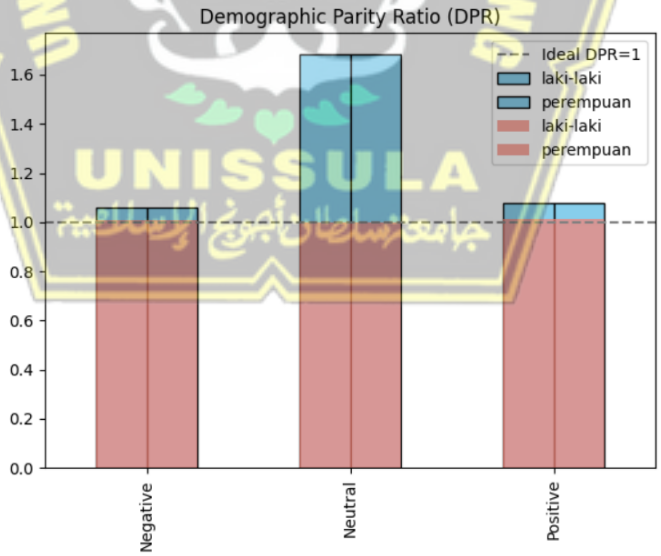
Secara keseluruhan, hasil mitigasi menunjukkan bahwa model dapat secara efektif mengurangi bias antar gender di seluruh kelas sentimen. Dengan nilai DPD yang hampir nol dan DPR yang hampir sama, perbedaan proporsi prediksi antara laki-laki dan perempuan sangat kecil. Oleh karena itu, mitigasi

berhasil meningkatkan keadilan (fairness) model. Ini terutama berlaku untuk kelas sentimen netral, yang sebelumnya menunjukkan bias yang paling kuat.



Gambar 4. 8 Visualisasi dpd sebelum dan setelah mitigasi

Gambar 4.8 merupakan visualisasi dari hasil sebelum dan sesudah mitigasi dari *demographic parity different.*



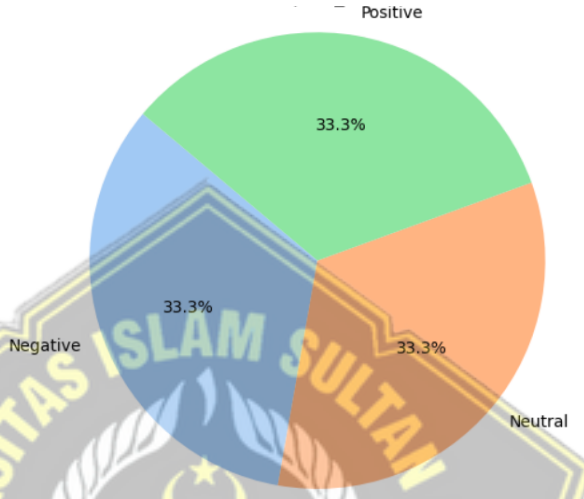
Gambar 4. 9 Visualisasi dpr sebelum dan setelah mitigasi

Gambar 4.9 merupakan visualisasi dari hasil sebelum dan sesudah mitigasi dari *demographic parity different.*

4.2.2 Model 2 (Model Dengan Data Yang Di Seimbangkan Keseluruhan)

1. Visualisasi distribusi data

   Pada model ini data murni yang sebelumnya di prosess kemudian di lakukan penyeimbangan/*balancing* yang bertujuan agar distribusi data latih seimbang dan sama rata.



Gambar 4. 10 Visualisasi distribusi data latih model 2

Pada gambar 4.10 menampilkan visualisasi distribusi data latih model 2 yang memiliki bobot sama rata yaitu 33% data negatif, 33% data positif, dan 33% data netral. Penyamarataan ini bertujuan untuk membuat model dengan data latih yang seimbang dan tidak condong pada satu kelas yang dominan.

2. *Finetuning* Model

   Dari hasil latih pada model ke 2 dengan masih menggunakan hal yang sama pada model ke 1 dimana menggunakan 3 *epoch*, model dimuat dengan *trainer API* dari *Huggingface* menggunakan konfigurasi *TrainingArguments* dengan menggunakan *epoch* 3 dalam pelatihannya, *batch size* 16 dan *learning rate* 2e-5. Dan mengasilkan metriks metriks yang dapat di evaluasi pada tahap selanjutnya.

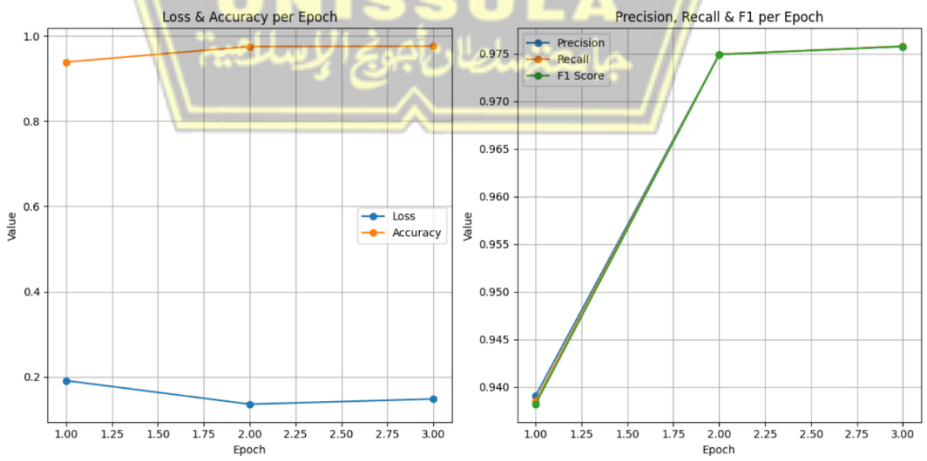
3. Evaluasi hasil pelatihan

   Setelah melakukan pelatihan pada model 2 tersebut dengan *finetuning indobert* lalu terdapat metriks berikut:

Tabel 4. 6 Hasil metriks pelatihan model 2

Validation_loss	***Accuracy***	***Precission***	***Recall***	***F1-Score***
0.147923	0.975758	0.975734	0.975758	0.975732

Berdasarkan tabel 4.6 Hasil pelatihan model menunjukkan metrik evaluasi yang sangat baik pada data validasi. Tingkat kesalahan prediksi model relatif rendah, dengan nilai kehilangan validasi 0,1479. Selain itu, nilai akurasi 97,57% menunjukkan bahwa model dapat memprediksi sebagian besar data validasi. Melihat metrik lainnya, kinerja model juga konsisten; nilai ketepatan 97,57 persen menunjukkan bahwa model dapat mengklasifikasikan data dengan benar tanpa banyak false positive. Namun, nilai recall 97,57 persen menunjukkan kemampuan model untuk mengumpulkan sebagian besar data penting tanpa mengalami banyak kehilangan kasus positif yang seharusnya terdeteksi. Nilai F1-Score sebesar 97,57%, yang menggabungkan precision dan recall, menunjukkan konsistensi ini. Secara keseluruhan, temuan ini menunjukkan bahwa model tidak hanya akurat tetapi juga seimbang dalam mengelola kesalahan klasifikasi, yang berarti bahwa itu bekerja dengan baik pada tahap validasi.

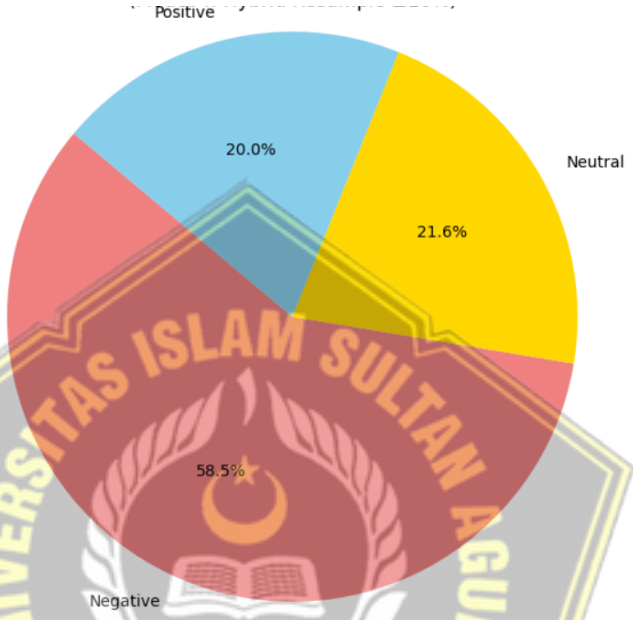


Gambar 4. 11 Visualisasi metriks hasil

Pada gambar 4.11 menampilkan visualisasi yang dihasilkan dari metriks pelatihan pada model 2.

4. Prediksi menggunakan model

Setelah dilakukanya prosess pelatihan pada data tersebut kemudian memuat kembali model yang sudah di latih tadi untuk dilakukan prediksi pada data test yang telah disiapkan sebelumnya. Dari hasil prediksi tersebut diperoleh sebgai berikut:



Gambar 4. 12 Visualisasi distribusi data test hasil prediksi

Dari gambar 4.12 menampilkan distribusi dari hasil prediksi menggunakan model yang telah dibangun dari model 2 dengan hasil 58% data negatif, 20% data positif dan 21 % data netral yang berhasil di prediksi berdasarkan model 2 yang telah di latih.

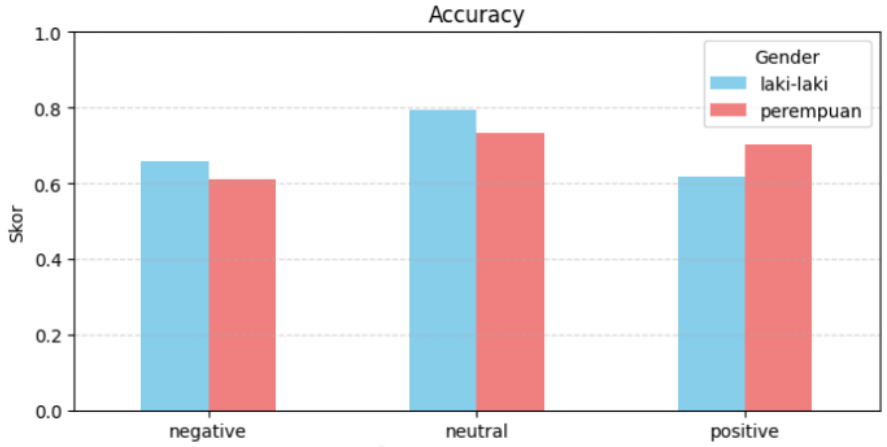
5. Evaluasi *Fairness* dasar dan mitigasi lanjutan *Fairness* dengan *Demographic parity*

Hasil evaluasi didasarkan pada output klasifikasi sentimen, yang akan dianalisis lebih lanjut untuk menemukan kemungkinan bias atau ketidakseimbangan prediksi antar *gender*. Hasil evaluasi *fairness* dasar terhadap model awal berikut:

Tabel 4. 7 Evaluasi fairness dasar

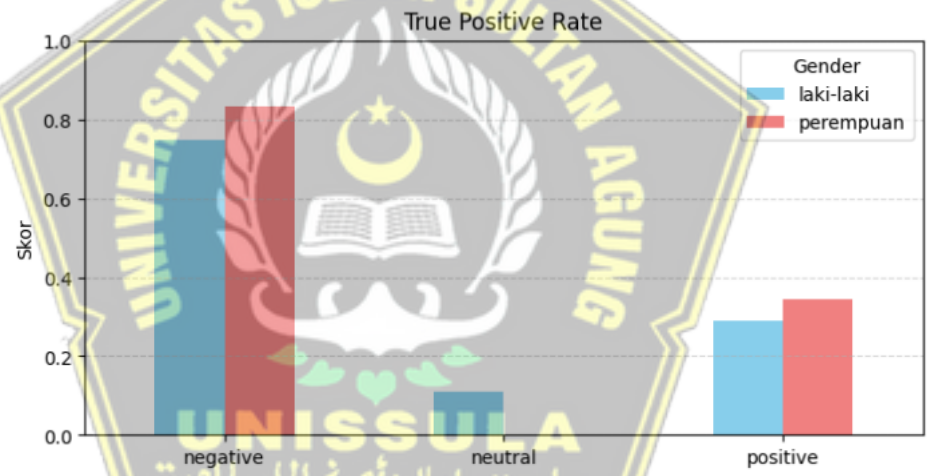
sentimen	Fairness metriks					
	Laki Laki			Perempuan		
	Accuracy	True positive	False positive	Accuracy	True positive	False positive
Positif	0.618	0.288	0.142	0.701	0.344	0.061
Negatif	0.658	0.747	0.450	0.610	0.831	0.581
Netral	0.794	0.108	0.184	0.733	0.0	0.148

Pada tabel 4.7 Hasil evaluasi fairness menunjukkan bahwa model masih mengalami ketidakadilan dalam performa prediksinya antar gender. Perempuan memperoleh akurasi (0,701) dan True Positive Rate (0,344) yang lebih tinggi dibandingkan laki-laki (0,618 dan 0,288). Selain itu, Rate False Positive perempuan (0,061) juga lebih rendah dibandingkan laki-laki. Ini menunjukkan adanya bias prediktif yang menguntungkan perempuan, di mana mereka lebih berpeluang mendapatkan prediksi yang benar pada kelas positif. Sebaliknya, pada kelas negatif, nilai TPR perempuan (0,831) lebih tinggi daripada nilai FPR laki-laki (0,747), dan nilai FPR perempuan (0,581) juga lebih tinggi daripada nilai laki-laki (0,450). Kondisi ini menunjukkan adanya bias prediktif yang cenderung merugikan perempuan, karena meskipun TPR perempuan lebih sering terdeteksi. Namun, bias yang paling signifikan ditemukan di kelas netral, di mana TPR perempuan sama sekali tidak terdeteksi (0,0), sedangkan TPR laki-laki masih mencapai 0,108. Ini menunjukkan bias kesempatan yang sama, karena perempuan tidak memiliki kesempatan yang sama dengan laki-laki untuk memprediksi sentimen netral dengan benar. Jadi, dapat disimpulkan bahwa model menunjukkan bias yang berbeda: perempuan di kelas positif diuntungkan, tetapi perempuan di kelas netral dan sebagian di kelas negatif diuntungkan.



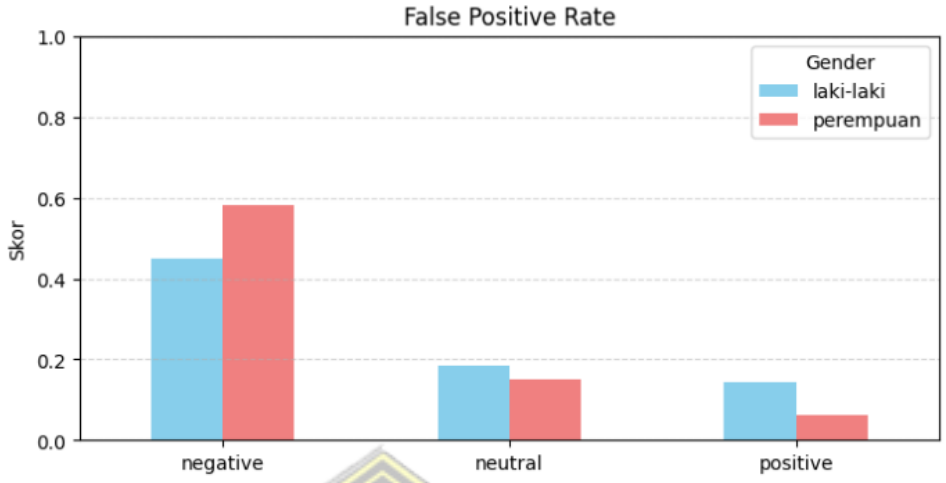
Gambar 4. 13 visualisasi metriks *Accuracy*

Meskipun tingkat akurasi masih sangat tinggi,masih terdapat bias antar kelompok memengaruhi kinerja model.



Gambar 4. 14 Visualisasi metriks TPR

Terdapat perbedaan besar antara kelompok, menurut visualisasi TPR. Ini menunjukkan bahwa kemampuan model untuk membuat prediksi positif tidak seragam.



Gambar 4. 15 visualisasi metriks FPR

Grafik FPR menunjukkan perbedaan antara kelompok, dengan kelompok tertentu cenderung lebih sering menerima prediksi yang salah.

Tabel 4. 8 Evaluasi fairness sebelum mitigasi

sentimen	Demographic parity					
	Laki Laki			Perempuan		
	Selection rate	Demographic parity difference	Demographic parity ratio	Selection rate	Demographic parity difference	Demographic parity ratio
Negatif	0.613	0.044	1.074	0.697	0.044	1.074
Netral	0.181	0.037	1.252	0.128	0.037	1.252
Positif	0.204	0.007	1.035	0.174	0.007	1.035

Pada tabel 4.8 Ada variasi ketidakseimbangan antar gender dalam masing-masing kelas sentimen, menurut hasil evaluasi fairness sebelum mitigasi berdasarkan metrik Demografi Paritas. Ada bias yang relatif kecil pada kelas negatif karena nilai Selection Rate (SR) laki-laki sebesar 0,613 lebih rendah daripada nilai SR perempuan sebesar 0,697. Nilai Demographic Parity Difference (DPD) sebesar 0,044 dan Demographic Parity Ratio (DPR) sebesar 1,074 keduanya cukup dekat dengan nilai ideal.

Kelas netral, di sisi lain, menunjukkan ketidakseimbangan yang lebih jelas. SR laki-laki (0,181) lebih tinggi daripada perempuan (0,128), dengan nilai DPD 0,037 dan DPR 1,252, dengan nilai yang lebih rendah dari 1 menunjukkan bahwa laki-laki lebih sering dianggap netral daripada perempuan, sehingga

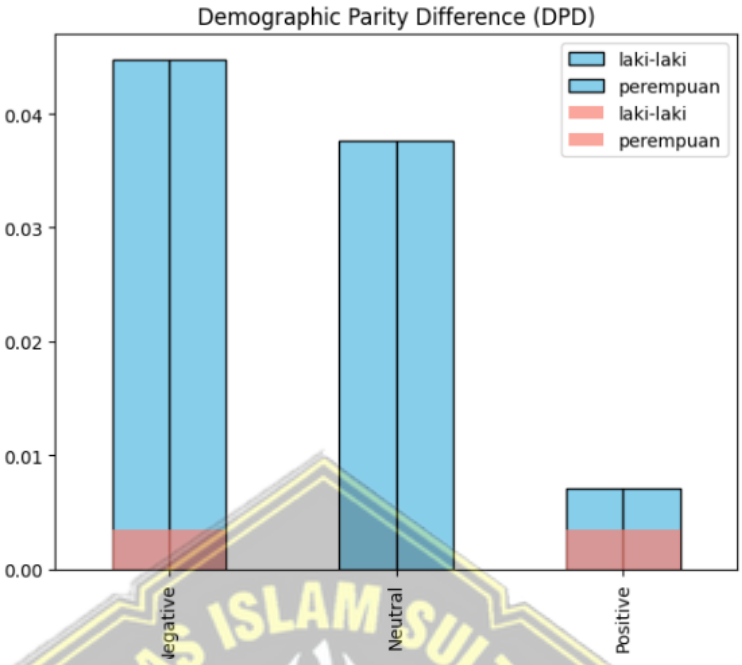
terjadi bias pada kelas ini. Di sisi lain, SR laki-laki (0,204) dan perempuan (0,174) cukup sebanding, dengan nilai DPD 0,007 dan DPR 1,035 yang hampir ideal. Ini menunjukkan bahwa prediksi positif didistribusikan secara seimbang antar gender. Dengan demikian, kelas netral memiliki bias paling dominan sebelum mitigasi, sedangkan kelas positif dan negatif menunjukkan distribusi gender yang lebih adil.

Tabel 4. 9 Evaluasi fairness setelah mitigasi

sentimen	Demographic parity					
	Laki Laki			Perempuan		
	Selection rate	Demographic parity difference	Demographic parity ratio	Selection rate	Demographic parity difference	Demographic parity ratio
Negatif	0.568	0.003	1.005	0.585	0.003	1.005
Netral	0.0	0.0	1.0	0.0	0.0	1.0
Positif	0.322	0.003	1.011	0.319	0.003	1.011

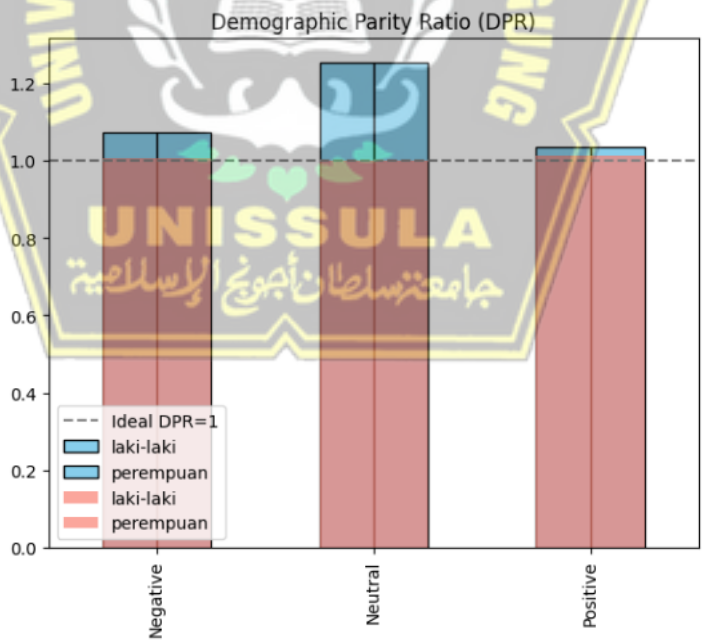
Hasil evaluasi fairness pasca mitigasi dengan metrik Paritas Demografi menunjukkan adanya perbaikan yang signifikan dalam distribusi prediksi antar gender. Nilai pilihan rasio (SR) laki-laki (0,568) dan perempuan (0,585) sangat sebanding di kelas negatif, dengan perbedaan demografis paritas (DPD) hanya 0,003 dan rasio paritas demografis (DPR) 1,005. Hasil ini menunjukkan bahwa bias telah diminimalkan karena model hampir tidak menunjukkan perbedaan perlakuan antara kedua gender dalam kelas negatif.

Hasil mitigasi menunjukkan keadaan yang hampir ideal pada kelas netral. SR laki-laki (0,0) dan perempuan (0,0) menghasilkan nilai DPD sebesar 0,0 dan DPR sebesar 1,0, yang menunjukkan bahwa tidak ada perbedaan dalam tingkat prediksi antara kedua jenis kelamin. Di kelas positif, distribusi prediksi laki-laki (0,322) dan perempuan (0,319) juga sebanding, dengan DPD sebesar 0,003 dan DPR sebesar 1,011, masing-masing tetap di bawah batas ideal. Secara keseluruhan, temuan ini menunjukkan bahwa strategi mitigasi berhasil mengurangi bias pada semua kelas sentimen; nilai DPD dan DPR hampir ideal, sehingga model dapat dianggap memiliki tingkat keadilan yang baik setelah mitigasi.



Gambar 4. 16 Viusalisasi DPD sebelum dan sesudah mitigasi

Gambar 4.16 merupakan visualisasi dari hasil sebelum dan sesudah mitigasi dari *demographic parity different.*



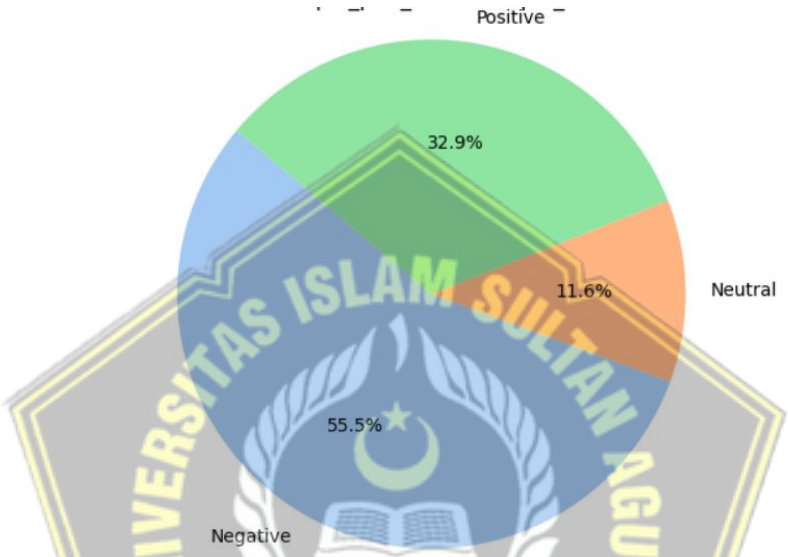
Gambar 4. 17 Viusalisasi DPR sebelum dan sesudah mitigasi

Gambar 4.16 dan 4.17 merupakan visualisasi dari hasil sebelum dan sesudah mitigasi dari *demographic parity different.*

4.2.3 Model 3 (Data Dengan Penyeimbangan Dan Pengurangan Sebanyak 10% Pada Kelas Mayoritas Dan Minoritas)

1. Visualisasi distribusi data

Model ke 3 ini dilakukan *balancing /* penyeimbangan pada kelas mayoritas dan kelas minoritas, dimana terdapat penambahan pada kelas minortias dan pengurangan pada kelas mayoritas sebanyak 10% pada keduanya.



Gambar 4. 18 Visualisasi distribusi data untuk model 3

Pada gambar 4.18 merupakan sebuah visualisasi data latih yang akan digunakan untuk membangun model 3 dimana distribusi tersebut sebanyak 55.5% merupakan data negatif kemudian 32.9% data positif dan 11.6% merupakan data netral. Bisa dibandingkan dari model ke 1 dimana sudah berkurang sebanyak 10% pada kelas mayoritas negatif dan bertambah pada kelas minoritas positif.

2. *Finetuning* Model

Untuk menjaga validitas perbandingan antar model, model ketiga dilatih dengan konfigurasi yang konsisten dengan model sebelumnya. Pelatih *Huggingface API* menjalankan proses pelatihan, dan parameter *TrainingArguments* mencakup jumlah *epoch* 3, ukuran *batch* 16, dan laju pembelajaran 2e-5. Parameter ini dipilih untuk memastikan stabilitas pelatihan dan mencegah data yang digunakan terlalu disesuaikan. Hasil pelatihan menghasilkan skor akurasi, presisi, recall, dan F1, yang akan digunakan sebagai

dasar untuk menganalisis kinerja model dan mengevaluasi aspek *fairness* pada tahap berikutnya.

3. Evaluasi hasil pelatihan

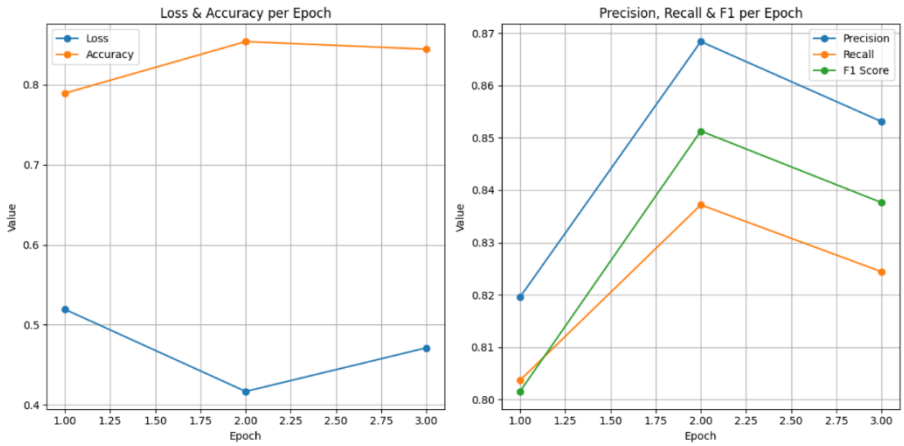
Setelah selesai melakukan pelatihan pada model ke 3 menghailkan beberapa parameter metriks seperti berikut:

Tabel 4. 10 Metriks evaluasi model 3

Validation_loss	***Accuracy***	***Precission***	***Recall***	***F1-Score***
0.416780	0.853890	0.868404	0.837186	0.851300

Pada tabel 4.8 Hasil evaluasi pelatihan model menunjukkan bahwa nilai kehilangan validasi berada pada 0.416780, menunjukkan bahwa tingkat kesalahan prediksi model pada data validasi masih relatif moderat, tetapi sudah cukup terkendali. Nilai akurasi sebesar 0.853890 menunjukkan bahwa sekitar 85% prediksi model sesuai dengan label sebenarnya, menunjukkan bahwa secara keseluruhan, data memiliki performa klasifikasi yang cukup baik.

Selain itu, metrik evaluasi tambahan menunjukkan kesesuaian kinerja model. Nilai precision sebesar 0.868404 menunjukkan kemampuan model untuk mengurangi jumlah kesalahan dalam memprediksi kelas positif dengan tingkat ketepatan yang tinggi, sementara nilai recall sebesar 0.837186 menunjukkan kemampuan model untuk mengumpulkan sebagian besar data yang relevan meskipun masih ada beberapa yang terlewat. Kombinasi keduanya menghasilkan F1-Score sebesar 0.851300, yang menunjukkan bahwa precision dan recall seimbang. Secara keseluruhan, temuan ini menunjukkan bahwa model memiliki kinerja klasifikasi yang baik dengan hasil yang seimbang antara akurasi, ketepatan, dan sensitivitas.

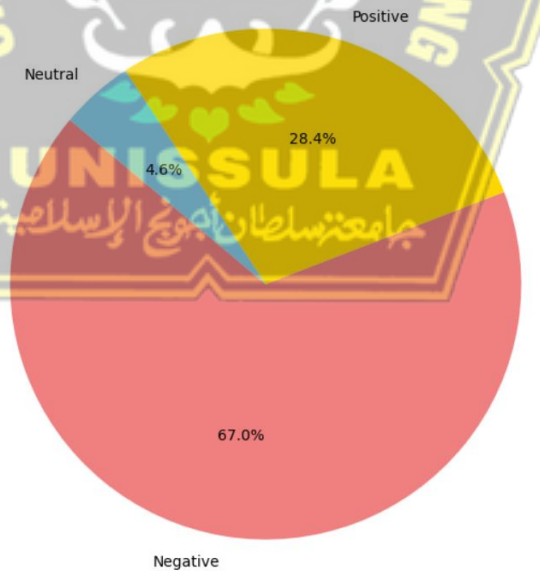


Gambar 4. 19 visualisasi metriks evaluasi pelatihan

Menampilkan visualisasi yang dihasilkan dari metriks pelatihan pada model 3.

4. Prediksi menggunakan model

Setelah melakukan pelatihan dan menyimpan hasil model tersebut kemudian model tersebut dimuat kembali untuk digunakan kembali dalam penggunaan prediksi pada data test yang sudah disiapkan dengan menghasilkan hasil prediksi sebagai berikut:



Gambar 4. 20 visualisasi distribusi hasil prediksi model 3

Pada gambar 4.20 hasil prediksi berdasarkan model 3 yang menghasilkan 67% berupa data negatif selanjutnya model mampu memprediksi 28.4% data positif dan yang terakhir sebanyak 4.6% berupa data netral.

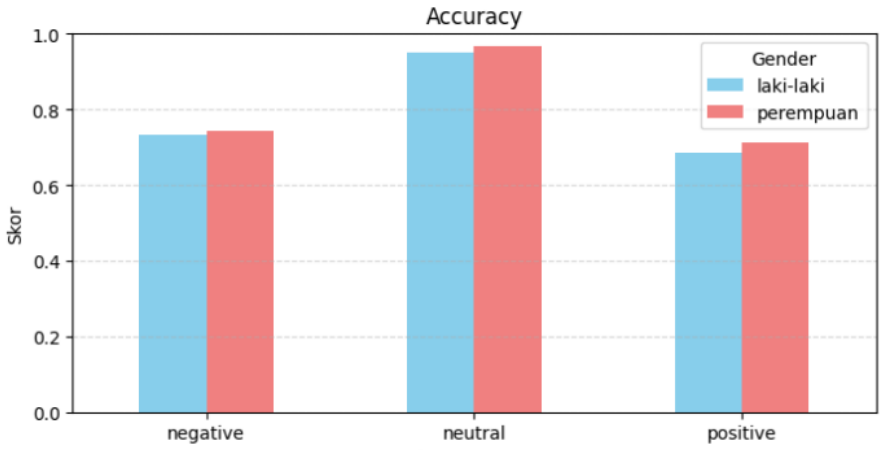
5. Evaluasi *Fairness* dasar dan mitigasi lanjutan *Fairness* dengan *Demographic parity*

Hasil evaluasi didasarkan pada output klasifikasi sentimen, yang akan dianalisis lebih lanjut untuk menemukan kemungkinan bias atau ketidakseimbangan prediksi antar *gender*. Hasil evaluasi *fairness* dasar terhadap model awal berikut:

Tabel 4. 11 Hasil evaluasi sebelum mitigasi

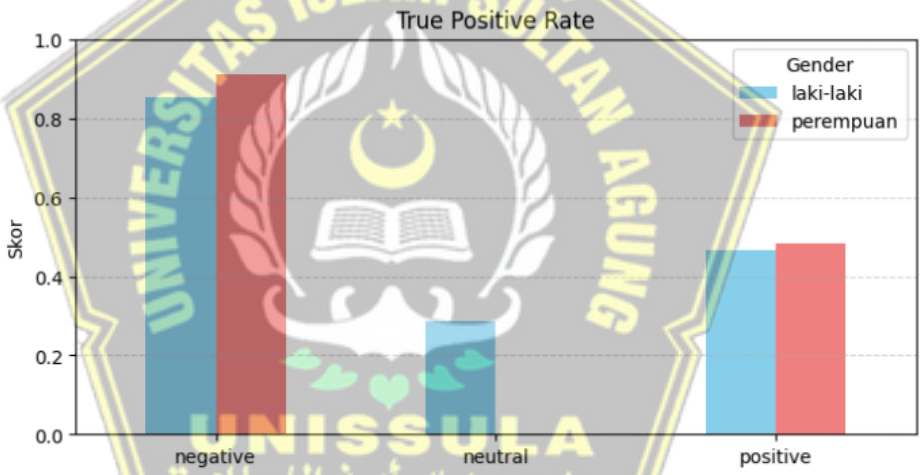
sentimen	**Fairness metriks**					
	Laki Laki			**Perempuan**		
	Accuracy	**True positive**	**False positive**	**Accuracy**	**True positive**	**False positive**
Positif	0.686	0.467	0.147	0.712	0.482	0.089
Negatif	0.732	0.853	0.423	0.744	0.910	0.448
Netral	0.948	0.285	0.047	0.968	0.0	0.031

Pada tabel 4.9 Nilai True Positive Rate (TPR) perempuan 0,482 sedikit lebih tinggi daripada laki-laki 0,467 pada sentimen positif namun, nilai False Positive Rate (FPR) laki-laki adalah 0,089 lebih rendah daripada perempuan. Ini menunjukkan bias prediksi; laki-laki lebih sering salah diklasifikasikan sebagai positif, sementara perempuan lebih jarang salah. TPR perempuan (0.910) lebih tinggi daripada laki-laki (0.853) pada sentimen negatif, dan FPR perempuan (0.448) juga lebih tinggi daripada laki-laki (0.423). Kondisi ini menunjukkan adanya trade-off bias: data negatif tentang perempuan lebih mudah ditemukan oleh model, tetapi prediksi negatif tentang perempuan juga lebih sering. Dengan kata lain, terjadi error bias yang tidak seimbang. Ketidaksesuaian sangat terlihat pada sentimen netral. Laki-laki masih memiliki TPR (0.285), tetapi model tidak dapat menemukan kasus netral untuk perempuan (0.0). Ini menunjukkan bias kesempatan yang serius, karena model tidak memberikan peluang kepada perempuan untuk diprediksi benar pada kelas netral.



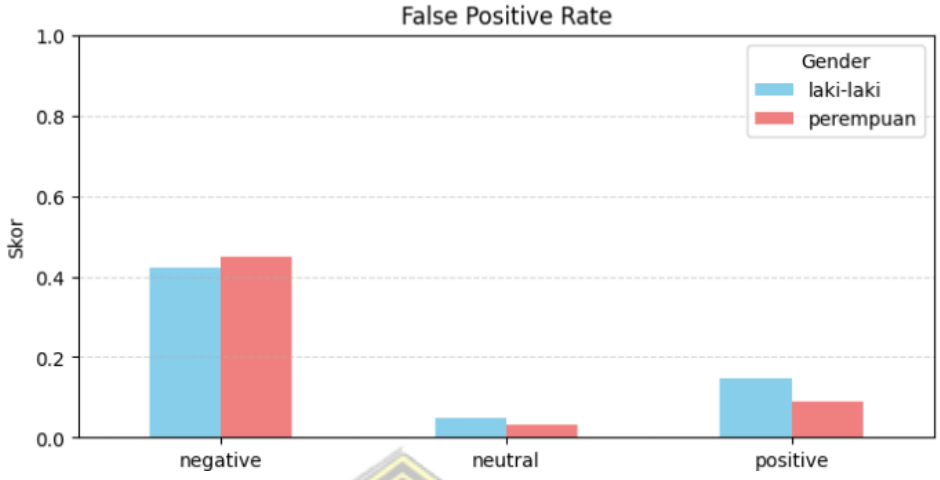
Gambar 4. 21 Visualisasi metriks akurasi

Meskipun tingkat akurasi masih sangat tinggi,masih terdapat bias antar kelompok memengaruhi kinerja model.



Gambar 4. 22 Visualisasi metriks TPR

Terdapat perbedaan besar antara kelompok, menurut visualisasi TPR. Ini menunjukkan bahwa kemampuan model untuk membuat prediksi positif tidak seragam.



Gambar 4. 23 Visualisasi metriks FPR

Dari gambar 4.21 sampai dengan 4.23 merupakan visualisasi dari hasil metriks seperti *accuracy, selection rate, true positive rate, false positive rate* dari evaluasi *fairness* dasar.

Tabel 4. 12 Hasil metriks sebelum mitigasi

sentimen	Demographic parity					
	Laki Laki			Perempuan		
	Selection rate	Demographic parity difference	Demographic parity ratio	Selection rate	Demographic parity difference	Demographic parity ratio
Negatif	0.665	0.031	1.046	0.696	0.031	1.046
Netral	0.048	0.016	1.527	0.031	0.016	1.527
Positif	0.285	0.014	1.053	0.271	0.014	1.053

Hasil evaluasi menunjukkan bahwa laki-laki dan perempuan memiliki perbedaan distribusi prediksi yang relatif kecil pada kelas negatif dan positif. Untuk kelas negatif, nilai Demographic Parity Difference (DPD) sebesar 0.031 dan untuk kelas positif sebesar 0.014, masing-masing. Namun, Demographic Parity Ratio (DPR) memiliki nilai yang hampir ideal, yaitu 1.046 dan 1.053. Hal ini menunjukkan bahwa model sudah cukup adil untuk memberikan peluang prediksi yang seimbang antara kedua kelompok dalam dua kelas tersebut dengan tingkat bias yang minimal.

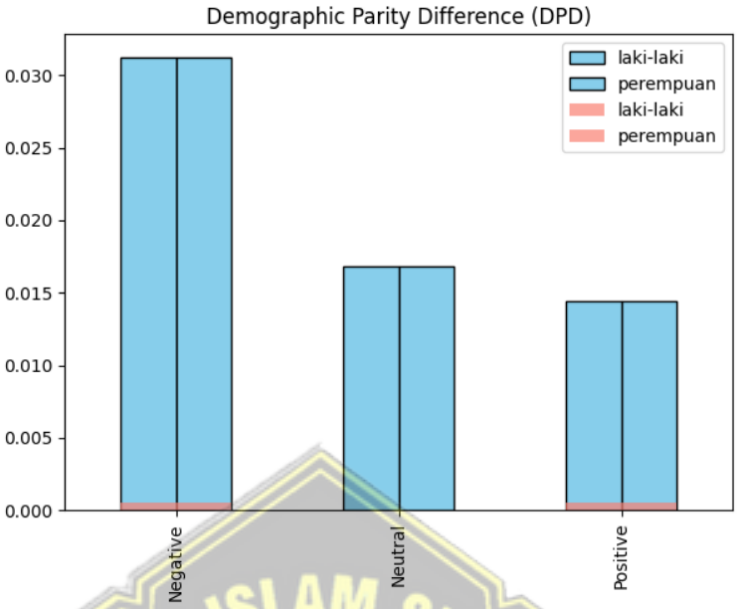
Namun, kelas netral menunjukkan ketimpangan yang lebih jelas. Jumlah pilihan (SR) laki-laki lebih tinggi (0.048) daripada perempuan (0.031), dan nilai DPR adalah 1.527, yang menunjukkan perbedaan yang signifikan dari nilai ideal. Ini menunjukkan bahwa data laki-laki lebih sering diklasifikasikan ke dalam kelas netral oleh model, sementara data perempuan cenderung kurang terwakili. Akibatnya, sebelum mitigasi, bias utama terletak pada sentimen netral. Sebaliknya, bias pada sentimen positif dan negatif sangat kecil.

Tabel 4. 13 Hasil metriks setelah mitigasi

sentimen	Demographic parity					
	Laki Laki			Perempuan		
	Selection rate	Demographic parity difference	Demographic parity ratio	Selection rate	Demographic parity difference	Demographic parity ratio
Negatif	0.548	0.000	1.000	0.547	0.000	1.000
Netral	0.0	0.0	1.0	0.0	0.0	1.0
Positif	0.410	0.000	1.001	0.409	0.000	1.001

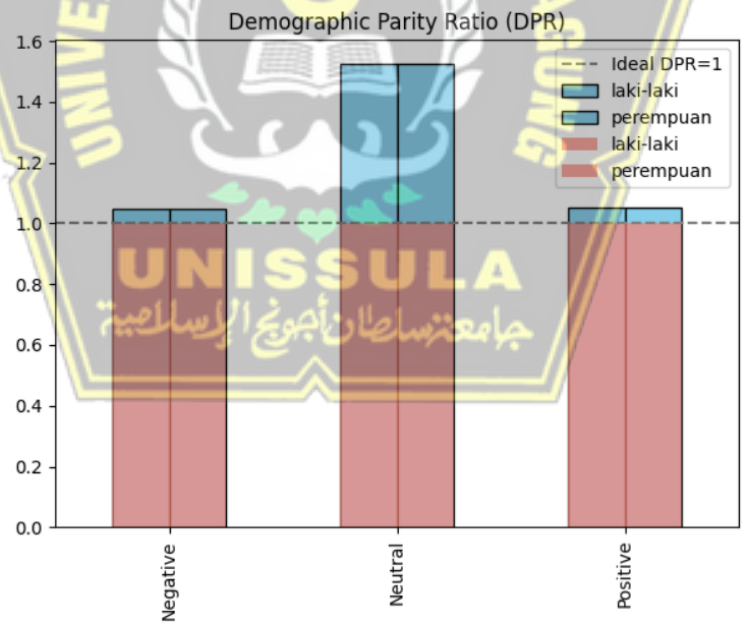
Hasil evaluasi setelah mitigasi menunjukkan bahwa model berhasil membuat kelompok gender lebih adil. Nilai Demographic Parity Difference (DPD) pada semua kelas sentimen (negatif, netral, dan positif) mendekati 0.000, dan nilai Demographic Parity Ratio (DPR) yang konsisten mendekati 1.000. Kondisi ini menunjukkan bahwa peluang prediksi untuk laki-laki dan perempuan hampir sama, sehingga tidak ada kelompok yang secara signifikan lebih diuntungkan atau dirugikan oleh model.

Khususnya, kelas netral, yang sebelumnya menunjukkan bias yang signifikan, sekarang seimbang, dengan SR laki-laki 0,0 dan perempuan 0,0, dengan DPD 0.0 dan DPR 1.0. Demikian pula, kelas positif dan negatif memiliki perbedaan yang sangat kecil dan tidak lagi menimbulkan bias yang signifikan, meskipun perbedaan pilihan rata-rata antar gender relatif kecil. Oleh karena itu, mitigasi terbukti efektif dalam mengurangi bias, yang membuat model lebih adil dan terbuka untuk membuat prediksi tentang semua kelas sentimen.



Gambar 4. 24 Viusalisasi DPD sebelum dan sesudah mitigasi

Gambar 4.24 merupakan visualisasi dari hasil sebelum dan sesudah mitigasi dari *demographic parity different.*



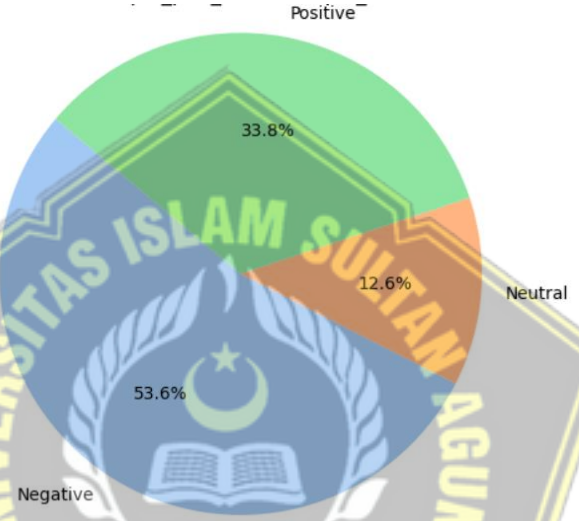
Gambar 4. 25 Viusalisasi DPR sebelum dan sesudah mitigasi

Gambar 4.25 dan 4.25 merupakan visualisasi dari hasil sebelum dan sesudah mitigasi dari *demographic parity different.*

4.2.4 Model 4 (Data Dengan Penyeimbangan Dan Pengurangan Sebanyak 20% Pada Kelas Mayoritas Dan Minoritas)

1. Visualisasi distribusi data

Model 4 ini dilakukan penyesuaian data latih yaitu dilakukanya *balancing* / penyeimbangan pada data latih dimana terdapat pengurangan dan penambahan pada data mayoritas dan data minoritas sebanyak 20% pada semua datanya. Berikut visualisasi distribusi pada data latih untuk model 4.



Gambar 4. 26 Distribusi data latih model 4

Berdasarkan pada gambar 4.26 didapatkan sebanyak 53.6% berupa data negatif kemudian sebanyak 33.8% berupa data positif dan yang terakhir sebanyak 12.6% merupakan data netral, untuk selanjutnya dilakukan prosess *finetuning* pada data latih tersebut.

2. *Finetuning* Model

Model keempat dibuat dengan menggunakan metode *resampling hybrid* dengan toleransi plus atau minus 10% pada distribusi label dan *gender* dalam data pelatihan. Mengurangi ketidakseimbangan distribusi, yang dapat memengaruhi kinerja dan kewajaran model, adalah tujuan. Pelatihan terus dilakukan melalui *API* Pelatihan *Huggingface*, dengan konfigurasi *TrainingArguments* yang sama, yaitu 3 *epoch*, ukuran *batch* 16, dan laju pembelajaran 2e-5. Untuk menjamin perbandingan hasil pelatihan antar model yang adil, parameter ini dipilih. Diharapkan model ini dapat menunjukkan

distribusi prediksi yang lebih baik untuk kelompok gender sensitif tanpa mengurangi akurasi umum.

3. Evaluasi hasil pelatihan

Hasil pelatihan model yang sudah dijalankan menghasilkan hasil hasil berikut:

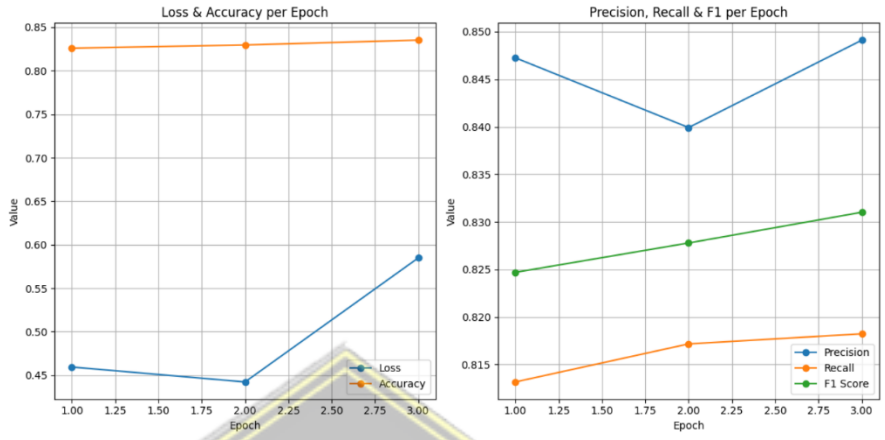
Tabel 4. 14 Metriks hasil pelatihan model 4

Validation_loss	***Accuracy***	***Precission***	***Recall***	***F1-Score***
0.584858	0.835185	0.849115	0.818227	0.831014

Hasil evaluasi pelatihan menunjukkan bahwa nilai kehilangan validasi sebesar 0.584858 masih tergolong moderat, yang menunjukkan bahwa model mengalami kesalahan prediksi sebagian, tetapi tidak terlalu banyak. Nilai akurasi sebesar 0.835185 menunjukkan bahwa model mampu membuat prediksi yang benar pada sekitar 83,5% data validasi. Oleh karena itu, performa klasifikasinya dianggap baik.

Selain itu, nilai precision sebesar 0.849115 menunjukkan kemampuan model untuk membuat prediksi yang tepat, terutama untuk meminimalkan kesalahan positif. Sementara itu, nilai recall sebesar 0.818227 menunjukkan bahwa model cukup mampu mendeteksi sebagian besar data yang relevan, meskipun masih ada beberapa kasus yang tidak teridentifikasi dengan benar. Kualitas F1 sebesar 0.831014 yang dihasilkan oleh kombinasi precision dan recall menunjukkan keseimbangan yang kuat antara ketepatan dan sensitivitas model. Secara keseluruhan, temuan ini menunjukkan bahwa model memiliki

kinerja klasifikasi yang stabil dan dapat digunakan pada tahap evaluasi yang lebih lanjut.

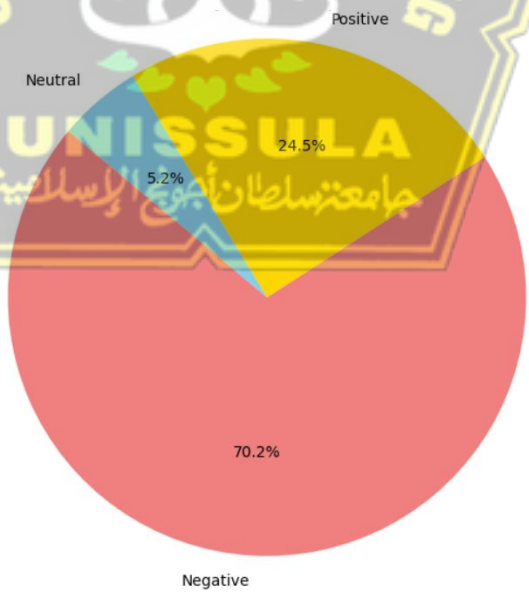


Gambar 4. 27 Visualisasi metriks pelatihan model 4

Bisa dilhat hasil visualisasi metriks yang dihasilkan dari pelatihan model 4 sebelumnya dan berhasil di visualisasikan dalam bentuk grafik.

4. Prediksi menggunakan model

Setelah dilakukannya pelatihan model kemudian tahap selanjutnya melakukan prediksi dengan model yang sudah dilatih dengan cara memuat ulang model tersebut yang telah disimpan.



Gambar 4. 28 visualisasi data hasil prediksi dengan model 4

Dari hasil prediksi tersebut dihasilkan berupa 70.2% berupa data negative kemudian 24.5% berupa data positif dan yang terakhir sebanyak 5.2% berupa data netral yang berhasil diprediksi dari model 4.

5. Evaluasi *Fairness* dasar dan mitigasi lanjutan *Fairness* dengan *Demographic parity*

Hasil evaluasi didasarkan pada output klasifikasi sentimen, yang akan dianalisis lebih lanjut untuk menemukan kemungkinan bias atau ketidakseimbangan prediksi antar *gender*. Hasil evaluasi *fairness* dasar terhadap model awal berikut:

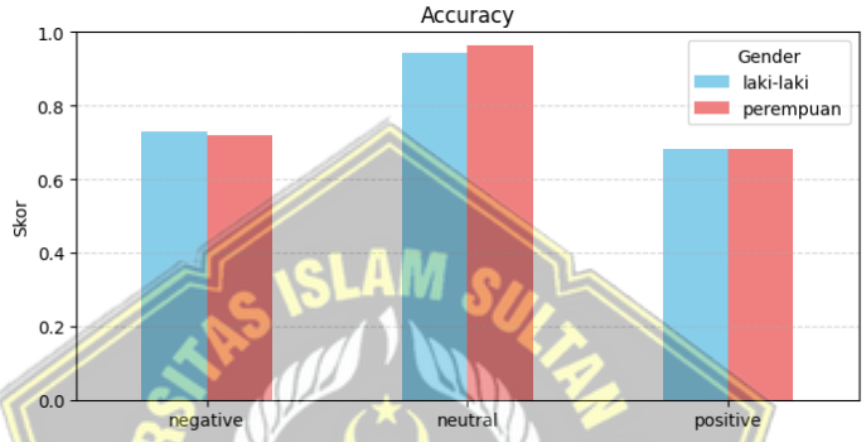
Tabel 4. 15 Hasil metriks fairness dasar

sentimen	**Fairness metriks**					
	Laki Laki			**Perempuan**		
	Accuracy	**True positive**	**False positive**	**Accuracy**	**True positive**	**False positive**
Positif	0.680	0.417	0.119	0.680	0.402	0.079
Negatif	0.731	0.880	0.461	0.718	0.920	0.517
Netral	0.944	0.428	0.052	0.962	0.0	0.037

Hasil evaluasi awal menunjukkan bahwa terdapat kecenderungan bias pada kelas sentimen tertentu. Pada sentimen positif, nilai True Positive Rate (TPR) laki-laki sedikit lebih tinggi daripada perempuan (0,402), tetapi nilai False Positive Rate (FPR) laki-laki juga sedikit lebih tinggi daripada perempuan (0,079). Ini menunjukkan adanya predictive bias, di mana model lebih sering mengklasifikasikan data laki-laki sebagai positif. Pada sentimen negatif, nilai TPR perempuan sedikit lebih tinggi (0,920) daripada laki-laki (0,880), dan FPR perempuan sedikit lebih rendah Kondisi ini menunjukkan adanya ketidakseimbangan kesalahan bias, karena kelompok perempuan memiliki sensitivitas yang lebih tinggi dan kesalahan prediksi yang lebih tinggi.

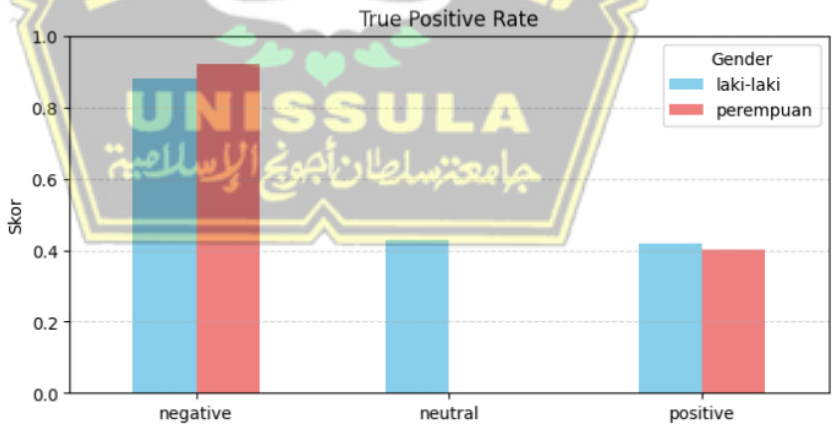
Sentimen netral menunjukkan bias yang lebih besar. Meskipun nilai TPR laki-laki 0,428 masih dapat diidentifikasi oleh model, nilai TPR perempuan 0,0, menunjukkan bahwa model tidak dapat mengidentifikasi data netral dalam

kelompok ini. Karena perempuan tidak memiliki kesempatan yang sama dengan laki-laki untuk membuat prediksi yang benar pada kelas netral, ini merupakan bentuk equal opportunity bias yang serius. Secara keseluruhan, dapat disimpulkan bahwa sebelum mitigasi, model menunjukkan kecenderungan bias yang berbeda di setiap kelas, mulai dari predictive bias, imbalanced error bias, dan equal opportunity bias.



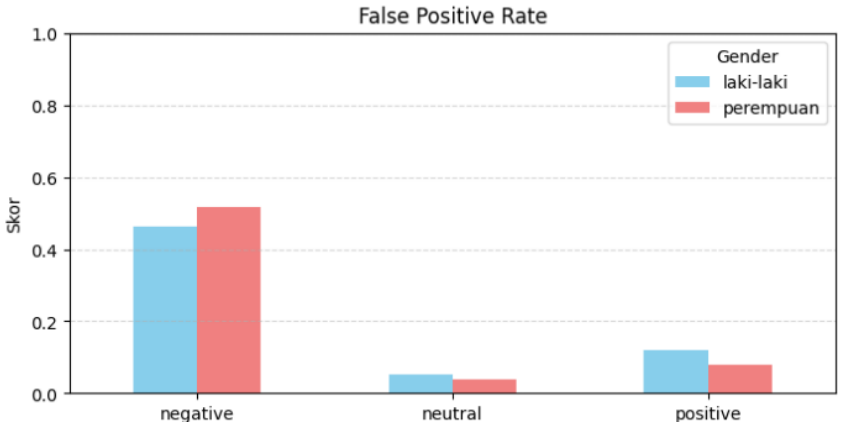
Gambar 4. 29 Visualisasi dari metriks akurasi

Meskipun tingkat akurasi masih sangat tinggi,masih terdapat bias antar kelompok memengaruhi kinerja model.



Gambar 4. 30 Visualisasi dari metriks TPR

Terdapat perbedaan besar antara kelompok, menurut visualisasi TPR. Ini menunjukkan bahwa kemampuan model untuk membuat prediksi positif tidak seragam.



Gambar 4. 31 Visualisasi dari metriks FPR

Visualisasi guna untuk menggambarkan hasil metriks dalam bentuk grafik visualisasi dan sebagai penggambaran metriks tersebut. Adapun berikut hasil dari mitigasi sebagai penanganan lebih lanjut dari hasil evaluasi sebelumnya.

Tabel 4. 16 Hasil metriks sebelum mitigasi

sentimen	Demographic parity					
	Laki Laki			Perempuan		
	Selection rate	Demographic parity difference	Demographic parity ratio	Selection rate	Demographic parity difference	Demographic parity ratio
Negatif	0.697	0.036	1.052	0.734	0.036	1.052
Netral	0.054	0.017	1.466	0.037	0.017	1.466
Positif	0.247	0.019	1.083	0.228	0.019	1.083

Hasil evaluasi menunjukkan bahwa perbedaan peluang prediksi antara laki-laki dan perempuan relatif kecil pada kelas negatif dan positif. Untuk kedua kelas tersebut, Demographic Parity Difference (DPD) hanya 0.036 pada kelas negatif dan 0.019 pada kelas positif, dengan Demographic Parity Ratio (DPR) masing-masing 1.052 dan 1.083, masing-masing. Nilai DPD hanya sekitar 0.036 pada kelas negatif dan DPR hanya 1.052, yang mendekati angka ideal 0, yang berarti bahwa model memberikan peluang prediksi yang hampir seimbang antar gender. Oleh karena itu, bias terhadap perasaan positif dan negatif dianggap rendah.

Namun, kelas netral menunjukkan ketidakseimbangan yang lebih jelas. Jumlah pilihan (SR) laki-laki lebih tinggi (0.054) daripada perempuan (0.037).

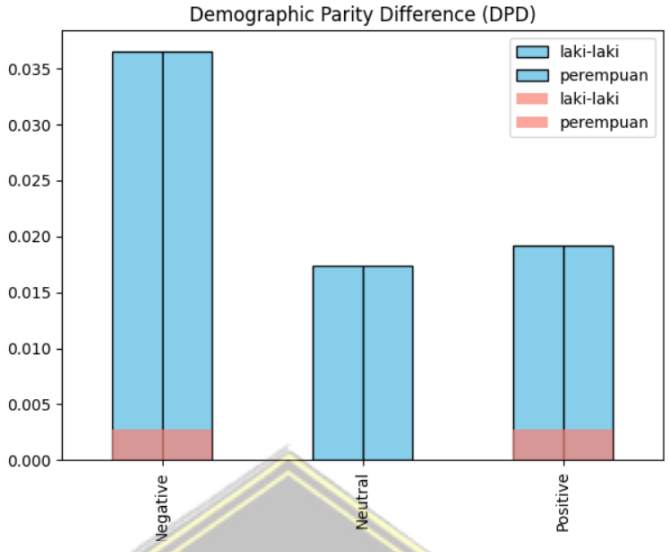
Nilai DPR adalah 1.466, jauh di bawah standar. Ini menunjukkan bahwa data dari kelompok laki-laki lebih sering diklasifikasikan ke dalam kategori netral daripada data dari kelompok perempuan. Kondisi ini menunjukkan bias representasi pada kelas netral, karena peluang perempuan untuk dianggap netral lebih rendah. Akibatnya, bias utama sebelum mitigasi terletak pada kelas netral, sementara bias minimal terlihat pada kelas negatif dan positif.

Tabel 4. 17 Hasil metriks setelah mitigasi

sentimen	Demographic parity					
	Laki Laki			Perempuan		
	Selection rate	Demographic parity difference	Demographic parity ratio	Selection rate	Demographic parity difference	Demographic parity ratio
Negatif	0.542	0.002	1.004	0.542	0.002	1.004
Netral	0.0	0.0	1.0	0.0	0.0	1.0
Positif	0.401	0.002	1.006	0.404	0.002	1.006

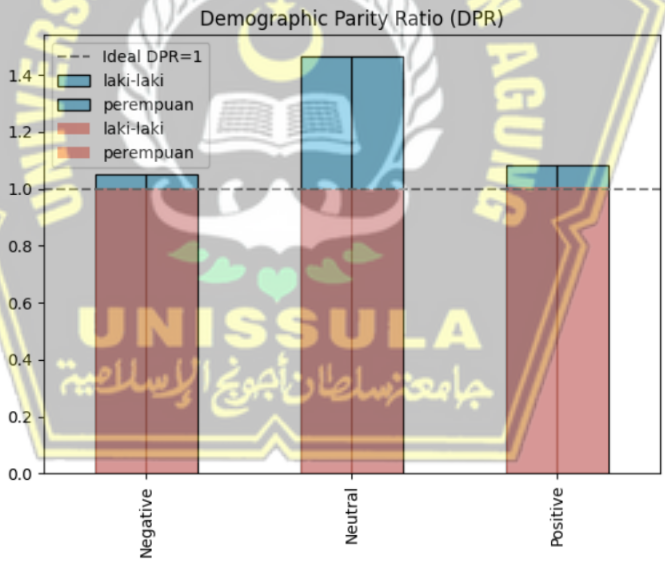
Hasil setelah mitigasi menunjukkan bahwa model menghasilkan keadilan antar gender yang lebih baik. Nilai Demographic Parity Difference (DPD) yang konsisten di seluruh kelas sentimen (negatif dan positif) adalah 0.002, dan nilai Demographic Parity Ratio (DPR) adalah tepat di antara 1.004. Kondisi ini menunjukkan bahwa peluang prediksi antara laki-laki dan perempuan sepenuhnya seimbang. Dengan demikian, model tidak memiliki kelompok yang lebih diuntungkan atau dirugikan.

Spesifik, kelas netral sebelumnya menunjukkan bias yang cukup signifikan dengan DPR lebih dari 1,5 setelah dikurangi, hasilnya menjadi seimbang sepenuhnya dengan DPD 0.0 dan DPR 1.0. Dalam kelas negatif dan positif, distribusi prediksi antara laki-laki dan perempuan telah sama, meskipun terdapat sedikit perbedaan dalam pilihan rasio (SR), tetapi nilai DPD nol menunjukkan bahwa perbedaan ini tidak lagi signifikan. Oleh karena itu, dapat disimpulkan bahwa proses mitigasi berhasil menghilangkan bias demografis dan meningkatkan keadilan prediksi model untuk semua kelas sentimen.



Gambar 4. 32 Viusalisasi DPD sebelum dan sesudah mitigasi

Gambar 4.32 merupakan visualisasi dari hasil sebelum dan sesudah mitigasi dari *demographic parity different.*



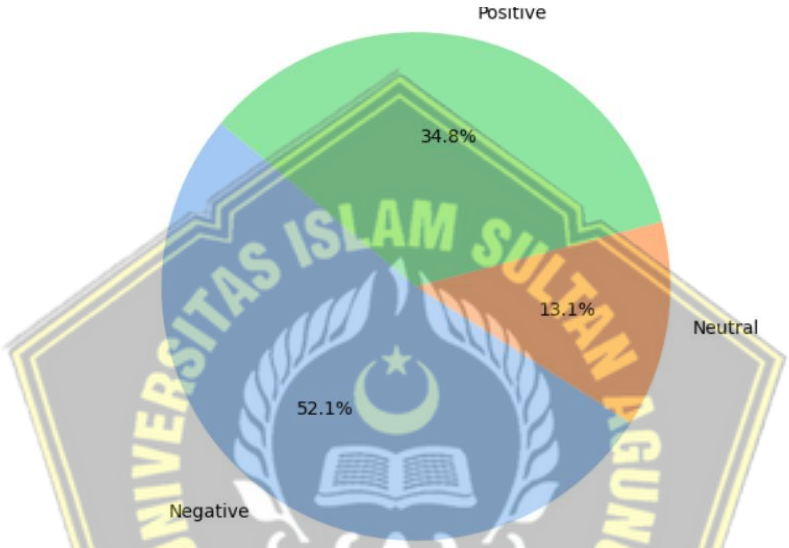
Gambar 4. 33 Viusalisasi DPR sebelum dan sesudah mitigasi

Gambar 4.33 merupakan visualisasi dari hasil sebelum dan sesudah mitigasi dari *demographic parity different.*

4.2.5 Model 5 (Data Dengan Penyeimbangan Dan Pengurangan Sebanyak 30% Pada Kelas Mayoritas Dan Minoritas)

1. Visualisasi distribusi data

Model 5 ini dilakukan penyesuaian data latih yaitu dilakukanya *balancing* / penyeimbangan pada data latih dimana terdapat pengurangan dan penambahan pada data mayoritas dan data minoritas sebanyak 30% pada semua datanya. Berikut visualisasi distribusi pada data latih untuk model 5.



Gambar 4. 34 Distribusi data latih model 4

Berdasarkan pada gambar 4.34 didapatkan sebanyak 52.1% berupa data negatif kemudian sebanyak 34.8% berupa data positif dan yang terakhir sebanyak 13.1% merupakan data netral, untuk selanjutnya dilakukan prosess *finetuning* pada data latih tersebut.

2. *Finetuning* Model

Untuk meningkatkan distribusi dan mengurangi kemungkinan bias, model kelima menggunakan strategi *resampling hybrid* ±30%. Prinsipnya serupa dengan model keempat, yaitu mengubah proporsi data pelatihan antara label dan *gender*. Metode dan konfigurasi pelatihan sama: *Trainer API* digunakan untuk memuat model dengan parameter *epoch* 3, ukuran *batch* 16, dan laju pembelajaran 2e-5. Diharapkan bahwa model kelima akan memberikan lebih banyak informasi tentang pengaruh tingkat penyesuaian distribusi terhadap metrik kinerja dan kewajaran. Itu juga akan menunjukkan seberapa besar

resampling dapat meningkatkan kesetaraan hasil model di seluruh kelompok *gender*.

3. Evaluasi hasil pelatihan

Setelah melakukan pelatihan pada model 5 tersebut dengan *finetuning indobert* lalu terdapat metriks berikut:

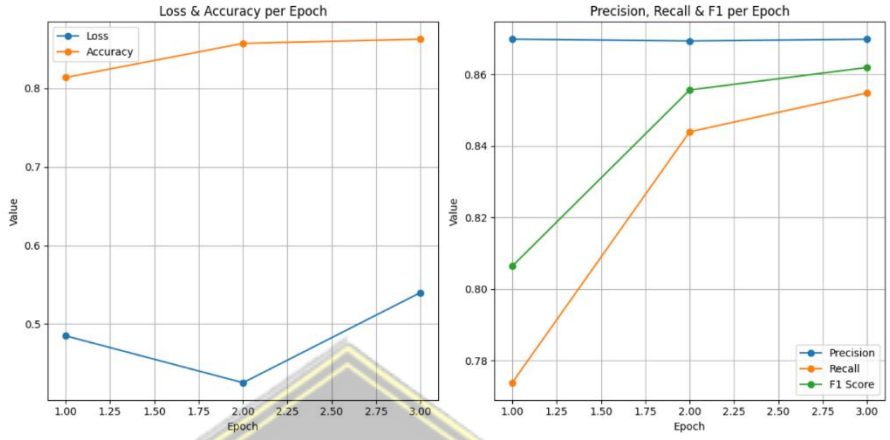
Tabel 4. 18 Metriks hasil pelatihan model 5

Validaton_loss	***Accuracy***	***Precission***	***Recall***	***F1-Score***
0.539984	0.862568	0.869814	0.854801	0.861918

Hasil pelatihan menunjukkan bahwa nilai kehilangan validasi sebesar 0.539984 masih berada pada tingkat moderat, yang menunjukkan bahwa tingkat kesalahan prediksi pada data validasi cukup terkendali dan tidak ada indikasi underfitting atau overfitting yang signifikan. Nilai akurasi sebesar 0.862568 menunjukkan bahwa model mampu memberikan prediksi yang benar pada sekitar 86% data, sehingga dapat dikategorikan dengan performa klasifikasi yang baik secara keseluruhan.

Selain itu, ada nilai precision sebesar 0,869814 yang menunjukkan bahwa model cukup konsisten dalam menghasilkan prediksi yang tepat dengan sedikit kesalahan positif. Di sisi lain, nilai recall sebesar 0.854801 menunjukkan bahwa model dapat mendeteksi sebagian besar data yang relevan, meskipun ada beberapa data yang belum ditemukan. Performa model ini cukup stabil dan dapat diandalkan, menurut F1-score sebesar 0.861918 yang dihasilkan oleh keseimbangan antara ketepatan dan recall ini. Secara keseluruhan, metrik

evaluasi ini menunjukkan bahwa model memiliki kinerja klasifikasi yang ideal dengan keseimbangan ketepatan dan sensitivitas.

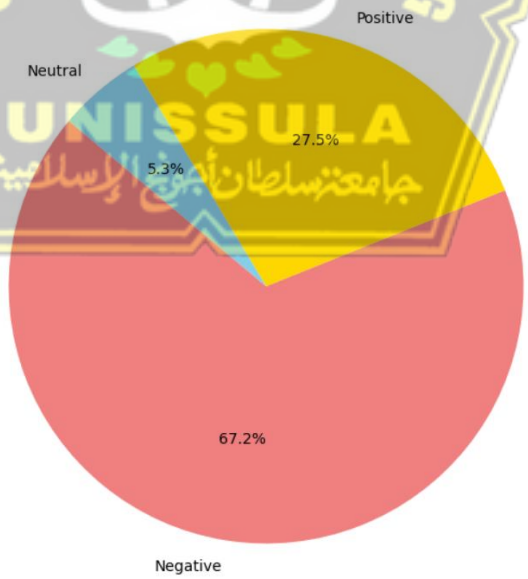


Gambar 4. 35 Visualisasi dari metriks pelatihan model *5*

Pada gambar 4.35 menampilkan visualisasi yang dihasilkan dari metriks pelatihan pada model 5.

4. Prediksi menggunakan model

Setelah dilakukannya pelatihan model kemudian tahap selanjutnya melakukan prediksi dengan model yang sudah dilatih dengan cara memuat ulang model tersebut yang telah disimpan.



Gambar 4. 36 Visualisasi data hasil prediksi dengan model 5

Dari hasil prediksi tersebut dihasilkan berupa 67.2% berupa data negative kemudian 27.5% berupa data positif dan yang terakhir sebanyak 5.3% berupa data netral yang berhasil diprediksi dari model 5.

5. Evaluasi *Fairness* dasar dan mitigasi lanjutan *Fairness* dengan *Demographic parity*

Hasil evaluasi didasarkan pada output klasifikasi sentimen, yang akan dianalisis lebih lanjut untuk menemukan kemungkinan bias atau ketidakseimbangan prediksi antar *gender*. Hasil evaluasi *fairness* dasar terhadap model awal berikut:

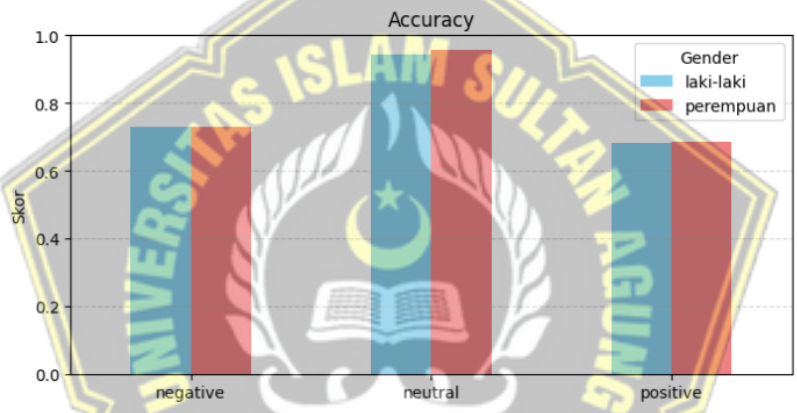
Tabel 4. 19 Hasil evaluasi metriks fairness dasar

sentimen	Fairness metriks					
	Laki Laki			Perempuan		
	Accuracy	True positive	False positive	Accuracy	True positive	False positive
Positif	0.680	0.452	0.146	0.702	0.45	0.111
Negatif	0.731	0.853	0.426	0.728	0.900	0.471
Netral	0.944	0.428	0.052	0.962	0.571	0.022

Berdasarkan hasil evaluasi awal terhadap metrik fairness, terlihat adanya perbedaan kinerja model dalam memprediksi sentimen berdasarkan kelompok gender. Pada sentimen positif, meskipun nilai *True Positive Rate* (TPR) relatif seimbang antara laki-laki (0.452) dan perempuan (0.450), terdapat ketimpangan pada *False Positive Rate* (FPR) di mana laki-laki (0.146) lebih tinggi dibandingkan perempuan (0.111). Kondisi ini menunjukkan adanya predictive bias, karena model cenderung lebih sering salah mengklasifikasikan data laki-laki sebagai sentimen positif. Sementara itu, pada sentimen negatif, perempuan memiliki TPR yang lebih tinggi (0.900) dibanding laki-laki (0.853), namun diikuti oleh FPR yang juga lebih besar (0.471 pada perempuan dibanding 0.426 pada laki-laki). Hal ini mengindikasikan adanya imbalanced error bias, yaitu

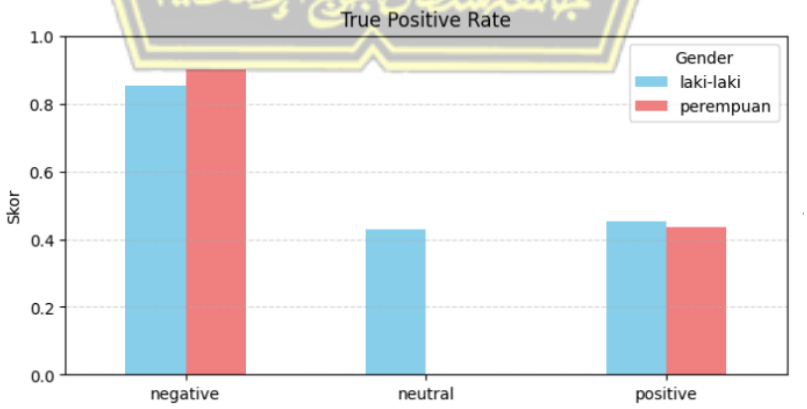
model lebih sensitif dalam mendeteksi sentimen negatif pada perempuan, tetapi di saat yang sama lebih banyak melakukan kesalahan pada kelompok tersebut.

Sentimen netral juga sangat berbeda. TPR perempuan lebih tinggi (0,571) dibandingkan laki-laki (0,428), dan FPR mereka lebih rendah (0,022) dibandingkan laki-laki. Kondisi ini menunjukkan bias, karena perempuan memiliki sensitivitas yang lebih tinggi dan kesalahan yang lebih rendah dalam memprediksi sentimen netral. Secara keseluruhan, temuan ini menunjukkan bahwa model belum sepenuhnya adil dalam memprediksi sentimen berdasarkan gender, karena laki-laki dan perempuan memiliki perbedaan dalam ketepatan dan kesalahan prediksi dalam masing-masing kategori sentimen.



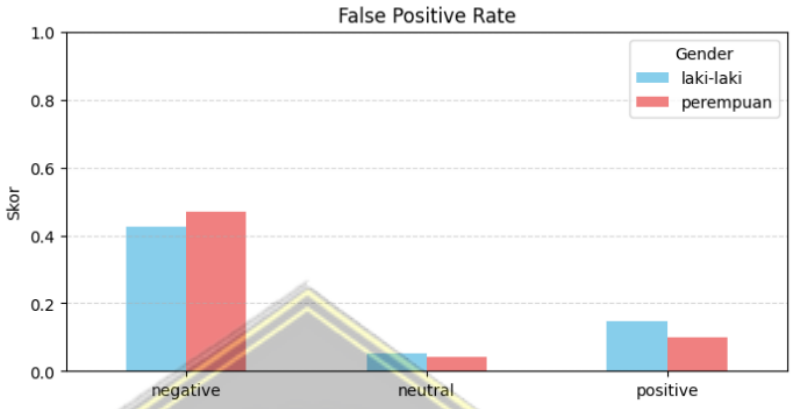
Gambar 4. 37 Visualisasi dari metriks akurasi

Meskipun tingkat akurasi masih sangat tinggi,masih terdapat bias antar kelompok memengaruhi kinerja model.



Gambar 4. 38 Visualisasi dari metriks TPR

Terdapat perbedaan besar antara kelompok, menurut visualisasi TPR. Ini menunjukkan bahwa kemampuan model untuk membuat prediksi positif tidak seragam.



Gambar 4. 39 Visualisasi dari metriks FPR

Visualisasi guna untuk menggambarkan hasil metriks dalam bentuk grafik visualisasi dan sebagai penggambaran metriks tersebut. Adapun berikut hasil dari mitigasi sebagai penanganan lebih lanjut dari hasil evaluasi sebelumnya.

Tabel 4. 20 Hasil metriks sebelum mitigasi

sentimen	Demographic parity					
	Laki Laki			Perempuan		
	Selection rate	Demographic parity difference	Demographic parity ratio	Selection rate	Demographic parity difference	Demographic parity ratio
Negatif	0.667	0.034	1.052	0.702	0.034	1.052
Netral	0.054	0.012	1.466	0.042	0.012	1.283
Positif	0.278	0.022	1.083	0.255	0.022	1.089

Hasil evaluasi fairness dengan metrik Demographic Parity sebelum mitigasi menunjukkan variasi dalam pilihan rasio (SR), perbedaan Demographic Parity (DPD), dan Demographic Parity Ratio (DPR) untuk setiap kategori sentimen. Untuk sentimen negatif, SR laki-laki cukup tinggi (0.667) dan perempuan (0.702). Namun, nilai DPD 0,034 dan DPR 1,052 menunjukkan perbedaan kecil dalam peluang prediksi sentimen negatif antar gender.

Indikasi bias yang lebih kuat ditemukan pada sentimen netral. Laki-laki memiliki SR lebih tinggi (0.054) daripada perempuan (0.042), dan DPD relatif

rendah (0.012), tetapi nilai DPR adalah 1.283, menunjukkan bahwa peluang prediksi sentimen netral untuk laki-laki lebih besar daripada perempuan, sehingga terdapat kecenderungan bias distribusi. Pada sentimen positif, SR laki-laki sedikit lebih tinggi (0.278) daripada perempuan (0.0255), dengan DPD sebesar 0.022 dan DPR sebesar 1.089. Ini menunjukkan bias kecil, di mana laki-laki lebih sering dianggap positif dibandingkan perempuan, meskipun perbedaan itu kecil.

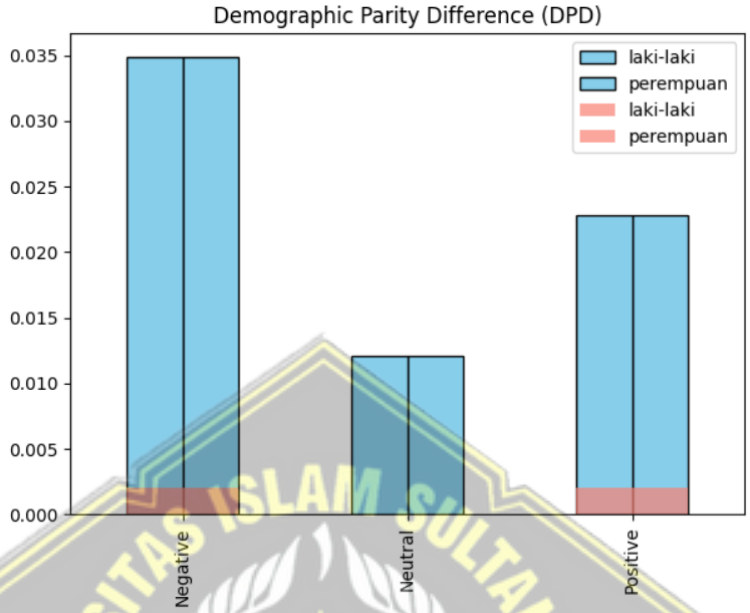
Tabel 4. 21 Hasil metriks sesudah mitigasi

sentimen	Demographic parity					
	Laki Laki			Perempuan		
	Selection rate	Demographic parity difference	Demographic parity ratio	Selection rate	Demographic parity difference	Demographic parity ratio
Negatif	0.584	0.001	1.002	0.590	0.001	1.002
Netral	0.0	0.0	1.0	0.0	0.0	1.0
Positif	0.378	0.001	1.003	0.377	0.001	1.003

Pada sentimen negatif, nilai pilihan rasio (SR) untuk laki-laki adalah 0.584 dan nilai pilihan rasio (SR) untuk perempuan adalah 0,590. Selain itu, nilai perbedaan rasio demografi (DPD) sangat rendah, yaitu 0,001 untuk laki-laki dan 0.001 untuk perempuan, dan Demographic Parity Ratio (DPR) adalah sekitar 1.002. Hal ini menunjukkan bahwa peluang untuk memprediksi perasaan buruk sama untuk kedua gender, sehingga bias hampir tidak ada. Untuk sentimen netral, baik laki-laki (SR = 0,0) maupun perempuan (SR = 0,0) memiliki nilai DPD = 0,0 dan DPR = 1,0, yang menunjukkan bahwa distribusi prediksi pada kelompok ini adil tanpa perbedaan gender. Dengan kata lain, mitigasi berhasil menghilangkan bias kelas netral.

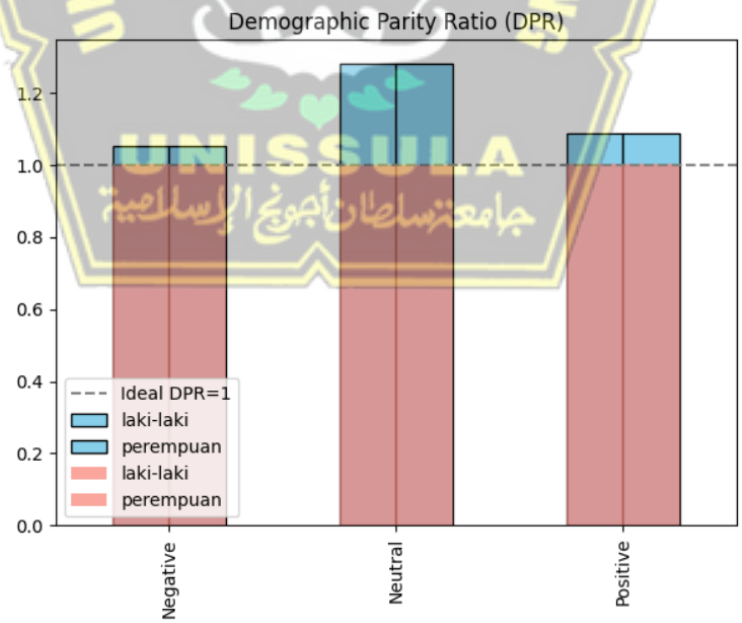
Namun, pada sentimen positif, nilai SR laki-laki (0,378) dan perempuan (0,377) hampir sama. DPR mendekati 1.003, dan DPD juga sangat kecil (0.001). Ini menunjukkan bahwa prediksi sentimen positif hampir setara antar gender, sehingga tidak ada indikasi bias yang signifikan. Secara keseluruhan, hasil setelah mitigasi menunjukkan distribusi prediksi antar gender menjadi lebih adil karena total nilai DPD mendekati nol dan nilai DPR mendekati 1. Akibatnya,

proses mitigasi berhasil mengurangi ketimpangan representasi sebelumnya, terutama pada sentimen positif dan netral.



Gambar 4. 40 salisasi DPD sebelum dan sesudah mitigasi

Gambar 4.40 merupakan visualisasi dari hasil sebelum dan sesudah mitigasi dari *demographic parity different.*



Gambar 4. 41 salisasi DPR sebelum dan sesudah mitigasi

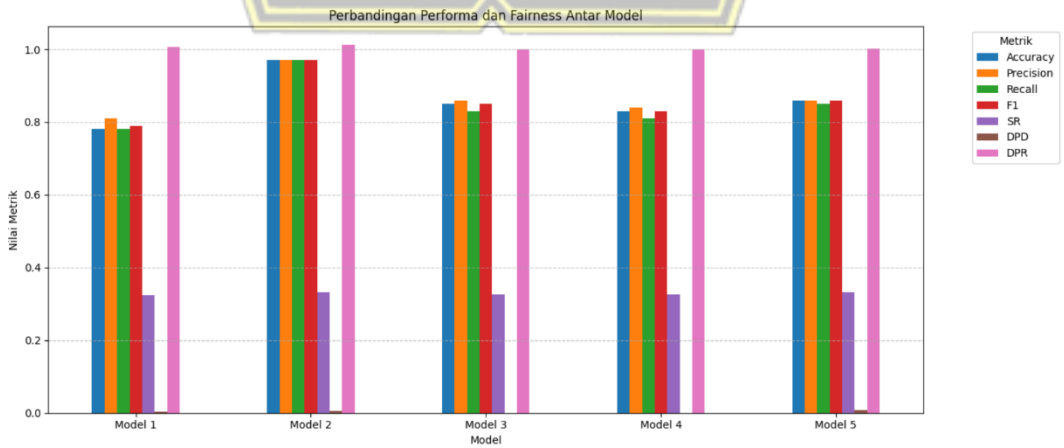
Gambar 4.41 merupakan visualisasi dari hasil sebelum dan sesudah mitigasi dari *demographic parity different.*

4.3 Perbandingan semua model

Tabel 4. 22 Perbandingan semua model

model	Wanita(%)	Pria(%)	Metriks evaluasi model				Demographic parity		
			accuracy	precision	recall	F1-score	Selection rate	Demographic parity difference	Demographic parity ratio
Model 1	25.3	74.6	0.78	0.81	0.78	0.79	0.323	0.003	1.006
Model 2	50	50	0.97	0.97	0.97	0.97	0.333	0.005	1.012
Model 3	34.5	65.5	0.85	0.86	0.83	0.85	0.327	0.000	1.000
Model 4	43.2	56.8	0.83	0.84	0.81	0.83	0.327	0.000	1.000
Model 5	51.5	48.5	0.86	0.86	0.85	0.86	0.332	0.007	1.003

Tabel perbandingan di atas menunjukkan bahwa ada perbedaan dalam performa dan fairness antar model. Model 3 menunjukkan hasil terbaik dari segi evaluasi performa. Model 1, yang memiliki distribusi data tidak seimbang (25,3% perempuan dan 74,6% laki-laki), menunjukkan kinerja yang lebih rendah (akurasi 0.78), yang menunjukkan potensi bias akibat ketidakseimbangan distribusi data, sementara Model 2, 4 dan 5 menunjukkan kinerja menengah yang lebih baik, menunjukkan proporsi data yang seimbang (50% laki-laki dan 50% perempuan), yang kemungkinan besar dipengaruhi oleh proporsi data yang seimbang.



Gambar 4. 42 Visualisasi hasil perbandingan semua model

Dari perspektif fairness (paritas demografi), sebagian besar model menghasilkan nilai DPD yang sangat kecil, bahkan hampir nol (Model 3 dan 4 = 0.000). Ini menunjukkan bahwa peluang prediksi positif antara pria dan wanita hampir sama. Selain itu, DPR (Demographic Parity Ratio) untuk seluruh model sangat dekat dengan 1 (di antara 1.000 dan 1.012), yang menunjukkan bahwa proporsi prediksi antar kelompok tetap stabil dan tidak terlalu bias. Tidak ada dominasi berlebihan dalam satu kelas, karena SR (pilihan rasio) tetap sama di seluruh model, berkisar antara 0.323 dan 0.333. kinerja. Secara keseluruhan, dapat disimpulkan bahwa mitigasi yang digunakan berhasil mengurangi bias gender (DPD rendah dan DPR ≈ 1). Namun, ini menghasilkan dampak negatif pada kinerja berbagai model, dengan Model 2 menunjukkan kinerja terbaik secara keseluruhan dari perspektif keadilan dan kinerja.

Tabel 4. 23 tabel trade off performa

model	Performa sebelum mitigasi fairness				Performa sesudah mitigasi fairness			
	accuracy	precision	recall	F1-score	accuracy	precision	recall	F1-score
Model 1	0.71	0.70	0.71	0.71	0.68	0.71	0.68	0.68
Model 2	0.62	0.60	0.62	0.60	0.56	0.65	0.56	0.57
Model 3	0.71	0.70	0.71	0.71	0.68	0.71	0.68	0.68
Model 4	0.69	0.68	0.69	0.68	0.67	0.70	0.67	0.66
Model 5	0.68	0.67	0.68	0.68	0.67	0.70	0.67	0.67

Seperti yang ditunjukkan dalam Tabel 4.23, hasil pengujian pada kelima model menunjukkan bahwa penerapan mitigasi fairness berdampak pada performa model. Secara umum, hampir semua model mengalami penurunan metrik akurasi, recall, dan F1-Score, meskipun terkadang ada peningkatan precision. Sebagai contoh, Model 1 mengalami penurunan metrik akurasi, recall, dan F1-Score dari 0,71 menjadi 0,68, atau sekitar 4,2%, sementara precision justru meningkat dari 0,70 menjadi 0,71, atau sekitar +1,4%. Model 3 juga mengalami penurunan metrik utama sebesar 4,2% tetapi peningkatan precision sebesar +1,4%. Model 2 menunjukkan trade-off yang lebih signifikan, di mana akurasi dan recall turun dari 0,62 menjadi 0,56 (-9,7%) serta F1 dari 0,60 menjadi 0,57 (-5%), tetapi precision

naik cukup tinggi dari 0,60 menjadi 0,65 (+8,3%). Pada Model 4 dan Model 5, penurunan performa terjadi dalam kisaran yang lebih kecil, yaitu sekitar -2,9% hingga -1,5% untuk akurasi, recall, dan F1-Score, sementara precision naik masing-masing sebesar +2,9% dan +4,5%.

Di antara kelima model yang diuji, Model 3 ditunjukkan sebagai yang paling stabil dan terbaik. Hasilnya di kedua tabel menunjukkan bahwa Model 3 memiliki nilai Demographic Parity Difference (DPD) 0.000 dan Demographic Parity Ratio (DPR) 1.000, yang menunjukkan bahwa tidak ada perbedaan signifikan dalam tingkat prediksi positif antara kelompok pria dan wanita. Nilai ini menunjukkan bahwa model telah mencapai tingkat keadilan yang tinggi tanpa bias demografis. Selain itu, dari segi performa, Model 3 tetap mempertahankan keseimbangan yang baik antara akurasi, presisi, recall, dan skor F1 baik sebelum maupun sesudah mitigasi fairness. Akurasi model hanya mengalami penurunan kecil dari 0.71 menjadi 0.68, yang wajar dan menunjukkan adanya perbedaan proporsional antara peningkatan fairness dan penurunan performa model. Oleh karena itu, Model 3 dapat dikatakan sebagai model terbaik karena berhasil mencapai keseimbangan terbaik antara kinerja (performance) dan tingkat keadilan (fairness). Model ini tidak menunjukkan bias yang signifikan antar gender dan menunjukkan penurunan performa yang masih di bawah batas normal.

Hasil ini menunjukkan adanya trade-off klasik dalam mitigasi fairness model yang lebih adil cenderung kehilangan akurasi dan konsistensi prediksi. Ini disebabkan oleh fakta bahwa algoritma mitigasi seperti ThresholdOptimizer mengubah ambang keputusan, juga dikenal sebagai ambang keputusan, untuk membuat distribusi prediksi lebih seimbang di antara kelompok sensitif. Akibatnya, walaupun kinerja klasifikasi tidak sebaik sebelum mitigasi, fairness model meningkat, yang diukur dengan Demographic Parity Difference (DPD) dan Demographic Parity Ratio (DPR). Menariknya, meskipun sebagian besar metrik menurun, empat dari lima model menunjukkan keakuratan yang lebih tinggi. Ini menunjukkan bahwa model menjadi lebih selektif saat melakukan prediksi, yang mengurangi kesalahan positif palsu.

Dengan demikian, dapat disimpulkan bahwa penerapan mitigasi fairness dalam penelitian ini berhasil menyeimbangkan proporsi prediksi antar gender, seperti yang ditunjukkan oleh nilai DPD yang hampir nol dan nilai DPR yang hampir 1. Namun, ada penurunan dalam kinerja model. Studi sebelumnya juga menemukan bahwa ada kompromi antara fairness dan performa model peningkatan fairness biasanya disertai dengan penurunan metrik evaluasi utama. Kondisi ini sejalan dengan temuan ini.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan beberapa poin penting sebagai berikut:

1. Kualitas model bergantung pada distribusi data. Ketidakseimbangan data gender dapat menyebabkan bias prediksi, sehingga intervensi seperti resampling diperlukan untuk memastikan representasi yang lebih adil.
2. Performa dan fairness memiliki trade-off. Model dengan akurasi 0.97% belum tentu paling adil, sementara model dengan nilai Demographic Parity 0.000 dapat menjaga keseimbangan prediksi antar gender meski akurasinya sedikit menurun dari 0.71 ke 0.68.
3. Fairness merupakan aspek penting dalam pemodelan AI. Akurasi tinggi tidak menjamin hasil prediksi yang adil, sehingga evaluasi dengan metrik seperti DPD dan DPR perlu dilakukan.
4. Mitigasi bias penting untuk meningkatkan keadilan dan kepercayaan publik, terutama dalam analisis opini masyarakat berbasis media sosial.

5.2 Saran

Berdasarkan penelitian yang telah dilakukan saran untuk penelitian penelitian selanjutnya adalah dengan menambah variasi dataset baik secara sentimen maupun gender supaya model memiliki kemampuan yang baik dalam mempelajari pola pola dari data yang beragam tersebut. Selain itu bisa mencoba metode mitigasi bias lainnya dan metode pembelajaran mesin lainnya seperti IndoBART, XLM-R dan jika terdapat bias dalam dataset agar lebih menggali lebih dalam performa yang di dapat dengan metode lainnya dalam upaya memitigasi bias tersebut. penelitian berikutnya dapat mengeksplorasi metode metode mitigasi penanganan keadilan pada model seperti metode pre-processing, in-processing guna menambah wawasan dalam penanganan keadilan model.

DAFTAR PUSTAKA

Agarwal, A. *et al.* (2020) "A reductions approach to fair classification," *35th International Conference on Machine Learning, ICML 2018*, 1, hal. 102–119.

Amatriain, X. *et al.* (2024) "Transformer models: an introduction and catalog." Tersedia pada: http://arxiv.org/abs/2302.07730.

Analysis, S. (1978) "Model Indo-BERT untuk Identifikasi Sentimen Kekerasan Verbal di Twitter," 18(x), hal. 583–593.

ANDRIYANI, W. *et al.* (2025) "Analisis Sentimen pada Ulasan Produk dengan SVM dan Word2Vec," *JIKO (Jurnal Informatika dan Komputer)*, 9(1), hal. 173. doi: 10.26798/jiko.v9i1.1498.

Anugerah Simanjuntak *et al.* (2024) "Research and Analysis of IndoBERT Hyperparameter Tuning in Fake News Detection," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, 13(1), hal. 60–67. doi: 10.22146/jnteti.v13i1.8532.

Ashktorab, Z. *et al.* (2023) "Fairness Evaluation in Text Classification: Machine Learning Practitioner Perspectives of Individual and Group Fairness," *Conference on Human Factors in Computing Systems - Proceedings*. doi: 10.1145/3544548.3581227.

Cahyawijaya, S. *et al.* (2021) "IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation," *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, hal. 8875–8898. doi: 10.18653/v1/2021.emnlp-main.699.

Cui, J. *et al.* (2023) *Survey on sentiment analysis: evolution of research methods and topics*, *Artificial Intelligence Review*. Springer Netherlands. doi: 10.1007/s10462-022-10386-z.

Czarnowska, P., Vyas, Y. dan Shah, K. (2021) "Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics," *Transactions of the Association for Computational Linguistics*, 9, hal. 1249–1267. doi: 10.1162/tacl_a_00425.

La Dahiri, M. (2020) “Pelaksanaan Tugas Dan Fungsi Pemerintahan Kecematan Ternate Utara Dalam Sistem Pelayanan Kepada Masyarakat Menurut Undang-Undang No. 9 Tahun 2015 Tentang Pemerintahan Daerah,” *Jurnal Ummu*, hal. 1059–1090.

Devlin, J. *et al.* (2019) “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm), hal. 4171–4186.

Dhendra dan Gayuh Utomo, V. (2025) “Benchmarking IndoBERT and Transformer Models for Sentiment Classification on Indonesian E-Government Service Reviews,” *Jurnal Transformatika*, 23(1), hal. 86–95. doi: 10.26623/transformatika.v23i1.12095.

Friedler, S. A. *et al.* (2020) “A comparative study of fairness-enhancing interventions in machine learning,” *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, hal. 329–338. doi: 10.1145/3287560.3287589.

Geni, L., Yulianti, E. dan Sensuse, D. I. (2024) “Analisis Sentimen Tweet Menjelang Pemilu 2024 di Indonesia Menggunakan Model Bahasa IndoBERT,” 9(3), hal. 746–757. doi: 10.26555/jiteki.v9i3.26490.

Hardt, M., Price, E. dan Srebro, N. (2016) “Equality of opportunity in supervised learning,” *Advances in Neural Information Processing Systems*, hal. 3323–3331.

Hidayat, M. N. dan Pramudita, R. (2024) “Analisis Sentimen Terhadap Pembelajaran Secara Daring Pasca Pandemi Covid-19 Menggunakan Metode IndoBERT,” *INFORMATION MANAGEMENT FOR EDUCATORS AND PROFESSIONALS : Journal of Information Management*, 8(2), hal. 161. doi: 10.51211/imbi.v8i2.2719.

Huang, X. (2022) “Easy Adaptation to Mitigate Gender Bias in Multilingual Text Classification,” *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Proceedings of the Conference, hal. 717–723. doi: 10.18653/v1/2022.naacl-main.52.

Kaur, H. dan Kaur Sandhu, D. (2023) "Evaluating the Effectiveness of the Proposed System Using F1 Score, Recall, Accuracy, Precision and Loss Metrics Compared to Prior Techniques," *International Journal of Communication Networks and Information Security*, 15(4), hal. 368–383. Tersedia pada: https://ijcnis.org.

Mahira Putri, Sutanto, T. E. dan Inna, S. (2023) "Studi Empiris Model BERT dan DistilBERT Analisis Sentimen pada Pemilihan Presiden Indonesia," *Indonesian Journal of Computer Science*, 12(5), hal. 2972–2980. doi: 10.33022/ijcs.v12i5.3445.

Makalew (2021) "Koordinasi Antara Pemerintah Dan Forum Kerukunan Umat Beragama (FKUB) Dalam Menciptakan Kerukunan Umat Beragama Di Kota Manado," *Jurnal governance*, 1(1), hal. 1–9. Tersedia pada: https://ejournal.unsrat.ac.id/index.php/governance/article/view/34304.

Merdiansah, R. dan Ali Ridha, A. (2024) "Sentiment Analysis of Indonesian X Users Regarding Electric Vehicles Using IndoBERT," *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI*, 7(1), hal. 221–228.

Mohiuddin, K. *et al.* (2023) "Retention Is All You Need," *International Conference on Information and Knowledge Management, Proceedings*, (Nips), hal. 4752–4758. doi: 10.1145/3583780.3615497.

BIN MUHAMMAD ALKATIRI, A., NADIAH, Z. dan NASUTION, A. N. S. (2020) "Opini Publik Terhadap Penerapan New Normal Di Media Sosial Twitter," *CoverAge: Journal of Strategic Communication*, 11(1), hal. 19–26. doi: 10.35814/coverage.v11i1.1728.

Nayla, A., Setianingsih, C. dan Dirgantoro, B. (2023) "Deteksi Hate Speech Pada Twitter," *eProceeding of Engineering*, 10(1), hal. 256.

Nia, Z. M. *et al.* (2023) "Twitter-based gender recognition using transformers," *Mathematical Biosciences and Engineering*, 20(9), hal. 15957–15977. doi: 10.3934/mbe.2023711.

Nurhasanah dan Zuriatin (2023) "Gender dan Kajian Teori Tentang Wanita,"

Edusociata Jurnal Pendidikan Sosiologi, 6(1), hal. 282–291. Tersedia pada: https://jurnal.stkipbima.ac.id/index.php/ES/article/view/1190/683.

Rangel, F. *et al.* (2018) "Overview of the 6th Author Profiling Task at PAN 2018: Multimodal gender identification in Twitter," *CEUR Workshop Proceedings*, 2125.

Resmadiktia, N. M., Utomo, Y. D. dan Aiman, L. M. (2023) "Pertanggungjawaban Pemerintah dalam Mewujudkan Good Governance sesuai Hukum Administrasi Negara," *Jurnal Ilmiah Wahana Pendidikan*, 9(11), hal. 685–697. Tersedia pada: https://jurnal.peneliti.net/index.php/JIWP/article/view/4394.

Saputra, A. C., Saragih, A. S. dan Ronaldo, D. (2025) "PREDIKSI EMOSI DALAM TEKS BAHASA INDONESIA," 19(1), hal. 1–15.

Saputra, F. T. *et al.* (2021) "Analisis Sentimen Bahasa Indonesia pada Twitter Menggunakan Struktur Tree Berbasis Leksikon," *Jurnal Teknologi Informasi dan Ilmu Komputer*, 8(1), hal. 135. doi: 10.25126/jtiik.0814133.

Sayarizki, P. dan Nurrahmi, H. (2024) "Implementation of IndoBERT for Sentiment Analysis of Indonesian Presidential Candidates," *Journal on Computing*, 9(2), hal. 61–72. doi: 10.34818/indojc.2024.9.2.934.

Siregar, A. (2022) "Efektivitas Penggunaan Media Sosial Sebagai Media Pendidikan," *EDU-RILIGIA: Jurnal Ilmu Pendidikan Islam dan Keagamaan*, 5(4), hal. 389–408. doi: 10.47006/er.v5i4.12936.

Takayasa, T. I. (2023) "The Implementation of Gender Mainstreaming Policy in Indonesian Local Government – The Case of Salatiga City 2017-2022," *JPW (Jurnal Politik Walisongo)*, 5(2), hal. 163–189. doi: 10.21580/jpw.v5i2.21121.

Tang, Z., Zhang, J. dan Zhang, K. (2023) "What-is and How-to for Fairness in Machine Learning: A Survey, Reflection, and Perspective," *ACM Computing Surveys*, 55(13 s). doi: 10.1145/3597199.

Utami, P. (2025) "Analisis Respons Publik di Media Sosial Terhadap Proses Legislasi RUU TNI Dalam Kerangka Demokrasi Deliberatif," *Communicology: Jurnal Ilmu Komunikasi*, 13(1), hal. 138–156. doi:

10.21009/comm.034.09.

Verma, S. dan Rubin, J. (2020) “Fairness definitions explained,” *Proceedings - International Conference on Software Engineering*, hal. 1–7. doi: 10.1145/3194770.3194776.

Wardani, B. S., Sa’adah, S. dan Nurjanah, D. (2023) “Evaluasi AI Bias and Fairness dalam Akuisisi Agen Penjualan Perbankan (Agen BRIlink-Bank Rakyat Indonesia),” *e-Proceeding of Engineering*, 10(6), hal. 5376–5384. Tersedia pada: https://foto.bisnis.com/.

Zeng, X., Dobriban, E. dan Cheng, G. (2022) “Fair Bayes-Optimal Classifiers Under Predictive Parity,” *Advances in Neural Information Processing Systems*, 35, hal. 1–24.