MUST: ***Journal of Mathematics Education, Science and Technology***
Vol. 10, No. 2, Desember 2025 Hal 1-17
DOI: http://doi.org/10.30651/must.v10i2.28707

AUGMENTASI GPT-4O DAN *FINE-TUNING INDOBERT* UNTUK ANALISIS SENTIMEN PUBLIK PADA ISU *RESHUFFLE* MENTERI KEUANGAN

Sabrina Adnin Kamila[1], Aisya Wina Wahda[2], Baiq Nina Febriati[3], Anwar Fitrianto[4], Rachmat Bintang Yudhianto[5]
[1, 2, 3, 4, 5]Departemen Statistika dan Sains Data, IPB University
sabrinaadnin@apps.ipb.ac.id[1], aisyawina@apps.ipb.ac.id[2],
baiqnina@apps.ipb.ac.id[3], anwarstat@gmail.com[4],
ydth_2000_rachmat@apps.ipb.ac.id[5]

Received 07 Oktober 2025; revised 03 Desember 2025; accepted 10 Desember 2025.

ABSTRAK

Reshuffle Menteri Keuangan pada tahun 2025 memicu perhatian publik luas dan menghasilkan dinamika opini di media sosial, khususnya X. Opini publik yang terekam dalam bentuk teks bersifat masif, *real-time*, dan tidak terstruktur, sehingga menghadirkan tantangan analisis karena penggunaan bahasa informal serta distribusi kelas sentimen yang tidak seimbang. Penelitian ini bertujuan untuk mengidentifikasi sentimen publik terkait *reshuffle* Menteri Keuangan dengan memanfaatkan integrasi GPT-4o dan *IndoBERT*. GPT-4o digunakan sebagai instrumen augmentasi data untuk memperkaya kelas minoritas, sedangkan *IndoBERT* berperan sebagai model klasifikasi sentimen yang dioptimalkan untuk bahasa Indonesia. Hasil penelitian menunjukkan bahwa pendekatan ini mampu meningkatkan kualitas representasi data dan stabilitas klasifikasi. Model *IndoBERT* yang dilatih dengan data hasil augmentasi mencapai akurasi 86% dan *macro-F1* sebesar 0,86, dengan performa terbaik pada kelas negatif (F1=0,88), disusul positif (F1=0,86) dan netral (F1=0,83). Temuan ini menegaskan bahwa integrasi GPT-4o dan *IndoBERT* efektif dalam mengatasi *imbalanced data* serta meningkatkan keandalan analisis sentimen berbahasa Indonesia. Penelitian ini tidak hanya memperkaya literatur analisis teks di Indonesia, tetapi juga memberikan kontribusi praktis bagi pembuat kebijakan dan media dalam memahami opini publik secara lebih proporsional.

Kata kunci: analisis sentimen, augmentasi data, GPT-4o, opini publik, IndoBERT.

ABSTRACT

The 2025 reshuffle of the Minister of Finance attracted wide public attention and sparked dynamic opinion on social media, particularly Twitter/X. Public opinion expressed in textual form is massive, real-time, and unstructured, posing analytical challenges due to the informal use of language and the imbalance of sentiment class distribution. This study aims to identify public

.

sentiment toward the reshuffle by integrating GPT-4o and IndoBERT. GPT-4o was employed as a data augmentation tool to enrich minority classes, while IndoBERT was applied as a sentiment classification model optimized for the Indonesian language. The findings indicate that this approach improves data representation and classification stability. The IndoBERT model trained on augmented data achieved 86% accuracy and a macro-F1 score of 0.86, with the best performance in the negative class (F1=0.88), followed by positive (F1=0.86) and neutral (F1=0.83). These results confirm that integrating GPT-4o and IndoBERT is effective in addressing imbalanced data and enhancing the reliability of Indonesian sentiment analysis. This research contributes to the growing body of literature on text analytics in Indonesia and provides practical implications for policymakers and the media in understanding public opinion more proportionally.

Keywords: sentiment analysis, public opinion, GPT-4o, data augmentation, IndoBERT

PENDAHULUAN

Reshuffle Menteri Keuangan pada tahun 2025 menarik perhatian publik karena posisi ini sangat strategis dalam menjaga stabilitas fiskal dan arah kebijakan ekonomi Indonesia. Pergantian pejabat di posisi tersebut tidak hanya memengaruhi mekanisme internal pemerintahan, tetapi juga memicu diskusi luas di kalangan masyarakat. Opini publik mencerminkan harapan, kritik, sekaligus apresiasi terhadap kebijakan ekonomi yang sedang dijalankan, sehingga analisis sentimen menjadi penting untuk memahami persepsi publik secara sistematis (Jim et al., 2024).

Perkembangan media sosial telah menggeser peran media konvensional seperti televisi, radio, dan surat kabar menuju ekosistem digital yang lebih terdesentralisasi dan interaktif (Canesta & Roestam, 2024). Media sosial X (Twitter) sebagai salah satu *platform* media sosial terbesar telah menjadi ruang diskursus publik yang dinamis. Karakteristiknya yang *real-time*, terbuka, dan interaktif membuat X merepresentasikan opini masyarakat dengan cepat dan masif (Rodríguez-Ibánez et al., 2023). Bahasa Indonesia di media sosial memiliki ciri khas yang berbeda dengan bahasa formal. Banyak pengguna memakai singkatan, bahasa gaul, emotikon, hingga sarkasme. Karakteristik ini menimbulkan tantangan dalam proses analisis sentimen dengan metode tradisional karena sulit memahami nuansa yang kompleks (Lin & Nuha, 2023).

Berbagai studi analisis sentimen berbahasa Indonesia sebelumnya banyak memanfaatkan metode klasik seperti *lexicon-based*, Naïve Bayes, dan *Support Vector Machine (SVM)*, serta arsitektur *deep learning* seperti *Long Short Term Memory* (LSTM) dan *Convolutional Neural Network* (CNN) terbukti efektif dalam berbagai konteks tetapi sering mengalami keterbatasan pada media sosial berbahasa Indonesia yang sangat informal. Penggunaan singkatan, *slang*, emotikon, sarkasme, dan struktur kalimat tidak baku cenderung mengacaukan fitur manual atau aturan leksikon (Ashar & Siahaan, 2024). Metode klasik umumnya bergantung pada *feature engineering* yang mudah terpengaruh istilah baru dan variasi bahasa lokal, sehingga model kurang stabil saat menghadapi teks baru. Perkembangan model *transformer* berbahasa Indonesia, khususnya *IndoBERT*, memiliki representasi kontekstual otomatis dan kemampuan adaptasi terhadap nuansa bahasa Indonesia yang beragam (Baharuddin & Naufal, 2023).

Penelitian Setyo Nugroho et al., (2021) melakukan analisis sentimen dalam ranah Bahasa Indonesia menggunakan model *BERT (fine-tune*) dan menghasilkan akurasi hingga 84% dengan *dataset* ulasan aplikasi Indonesia. Penelitian Aras et al., (2024) juga menunjukkan bahwa *IndoBERT* dapat diterapkan pada domain *e-commerce* Indonesia untuk mengklasifikasikan sentimen ulasan pengguna dengan hasil yang kompetitif. Selain tantangan linguistik, isu *imbalanced data* menjadi masalah mendasar dalam analisis sentimen. Model yang dilatih pada data yang tidak seimbang cenderung bias ke kelas mayoritas karena frekuensi yang lebih tinggi memperkuat sinyal kelas tersebut dalam proses optimisasi (Carvalho et al., 2025). Beberapa metode seperti *oversampling, undersampling,* maupun *Synthetic Minority Oversampling Technique* (SMOTE) telah digunakan untuk mengatasi masalah ini, namun pendekatan tersebut kurang optimal untuk teks karena berpotensi menimbulkan *overfitting* atau kehilangan konteks linguistik (Alkhawaldeh et al., 2023).

Perkembangan *Large Language Model* (LLM) seperti GPT-4o menawarkan solusi baru untuk augmentasi data. Berbeda dengan *oversampling* klasik, GPT-4o mampu menghasilkan variasi kalimat sintesis yang tetap konsisten dengan polaritas aslinya, namun lebih beragam secara linguistik. Pendekatan ini sejalan dengan penelitian Wei dan Zou (2019) yang menekankan pentingnya augmentasi teks

berbasis variasi linguistik untuk meningkatkan generalisasi model. Kualitas augmentasi dapat dievaluasi menggunakan *novelty* untuk mengukur kebaruan data sintesis dan *diversity* untuk menilai keragaman kosakata. *Novelty* merupakan indikator utama untuk mengevaluasi kualitas augmentasi teks dengan mengukur sejauh mana data sintesis berbeda dari data asli. Konsep ini awalnya diperkenalkan dalam bidang *information retrieval* oleh Carbonell (1998) dan kemudian diadopsi dalam evaluasi teks generatif *modern* oleh Duffy et al. (2025).

Berdasarkan *gap* tersebut, enelitian ini mengintegrasikan GPT-4o untuk augmentasi data dan *IndoBERT* untuk klasifikasi sentimen. GPT-4o dipilih karena kemampuannya menghasilkan teks sintesis yang konsisten dengan polaritas asli sekaligus lebih bervariasi secara linguistik (Suhaeni & Yong, 2023). Evaluasi performa model dilakukan menggunakan metrik akurasi dan *F1-score* untuk menilai ketepatan dan keadilan performa antar kelas, sementara kualitas augmentasi teks ditinjau dari aspek kebaruan dan keragaman bahasa.

Tujuan penelitian ini adalah: (1) mengidentifikasi sentimen publik terhadap *reshuffle* Menteri Keuangan menggunakan data X, (2) menyeimbangkan distribusi kelas sentimen melalui augmentasi berbasis GPT-4o yang dievaluasi dengan indikator *novelty* dan *diversity*, serta (3) mengevaluasi performa *IndoBERT* menggunakan metrik akurasi dan *F1-score*. Hipotesis yang diajukan adalah bahwa kombinasi augmentasi GPT-4o dengan *fine-tuning IndoBERT* mampu menghasilkan model analisis sentimen yang representatif, stabil, dan adil pada data media sosial berbahasa Indonesia.

METODE PENELITIAN

Subjek penelitian ini berupa data teks dari media sosial X (Twitter) terkait *reshuffle* Menteri Keuangan di Indonesia. Data dikumpulkan menggunakan metode *web scraping* dengan kata kunci Sri Mulyani, Purbaya, dan Menkeu pada periode 8-16 September 2025 yang bertepatan dengan momentum pengumuman diskusi publik terkait *reshuffle* tersebut. Data yang berhasil dikumpulkan kemudian diperiksa kualitasnya termasuk penghapusan cuitan duplikat untuk memastikan keunikan informasi yang dianalisis. Setelah proses pembersihan, jumlah data yang digunakan dalam penelitian ini adalah sebanyak 4.293 cuitan. Seluruh proses

pengolahan dan analisis dalam penelitian ini dilakukan dengan perangkat komputasi berbasis Python.

Variabel utama dalam penelitian ini adalah teks cuitan yang dianalisis berdasarkan sentimen dengan kategori positif, negatif, dan netral. Variabel tambahan meliputi informasi waktu publikasi cuitan (tanggal dan jam) serta keterlibatan pengguna seperti jumlah suka (*favorite_count*), *retweet,* dan jumlah balasan. Variabel target berupa label sentimen diperoleh melalui bantuan model GPT-4o yang digunakan untuk pelabelan awal dan diverifikasi secara manual guna memastikan reliabilitas data.

Instrumen penelitian meliputi perangkat lunak Python beserta pustaka pendukung seperti *transformers, scikit-learn, matplotlib, seaborn,* dan *wordcloud.* Model GPT-4o digunakan untuk *labeling* dan augmentasi teks, sedangkan model IndoBERT digunakan untuk proses *fine-tuning* dalam analisis sentimen.

Langkah-langkah penelitian dimulai dari pengumpulan data dengan *web scraping,* kemudian dilakukan proses pelabelan sentimen menggunakan GPT-4o yang diverifikasi manual. Selanjutnya dilakukan pra-pemrosesan data yang mencakup *case folding,* penghapusan URL, penghapusan *mention,* penggantian *hashtag* menjadi kata aslinya, pembersihan tanda baca, normalisasi spasi ganda, penghapusan *stopwords,* serta normalisasi teks. Setelah pra-pemrosesan, dilakukan analisis eksploratif awal untuk memahami karakteristik data.

Analisis distribusi sentimen dilakukan untuk mengidentifikasi ketidakseimbangan data. Ketidakseimbangan tersebut diatasi dengan melakukan augmentasi teks menggunakan GPT-4o pada kelas minoritas (negatif dan netral). Kualitas data sintesis kemudian dievaluasi melalui ukuran *novelty* (kebaruan) dan *diversity* (keragaman).

Nilai *novelty* tinggi yang mendekati 1 menandakan bahwa teks hasil augmentasi benar-benar baru dan bukan sekadar duplikasi dari data yang sudah ada. *Novelty* dihitung dengan membandingkan proporsi data augmentasi yang tidak terdapat pada data asli (Duffy et al., 2025):

$$Novelty = 1 - \frac{|D_{aug} \cap D_{ori}|}{D_{aug}}$$

Diversity berfokus pada keragaman kosakata yang muncul dalam data sintesis. Ukuran ini umumnya dievaluasi menggunakan *distinct-2* yang menghitung proporsi bigram unik terhadap total bigram dalam korpus (Duffy et al., 2025). Nilai *diversity* yang tinggi menunjukkan bahwa kalimat hasil augmentasi mengandung variasi kata dan kombinasi frasa yang lebih luas. *Diversity* dihitung dengan rumus (Papineni et al., 2001):

$$Diversity_{distinct-2} = \frac{Jumlah\ bigram\ unik}{Total\ bigram}$$

Teknik analisis data dalam penelitian ini terdiri dari dua tahap utama. Pertama, analisis deskriptif dengan visualisasi distribusi kata, waktu, dan kelas sentimen untuk memberikan gambaran umum data. Kedua, analisis prediktif menggunakan *fine-tuning* IndoBERT dengan *Trainer API* dari *Hugging Face* menggunakan parameter pelatihan 3 *epoch, batch size 16, learning rate 2e-5,* dan *weight decay* 0,01.

Evaluasi model klasifikasi dilakukan dengan akurasi dan *F1-score.* Akurasi mengukur proporsi prediksi yang benar terhadap seluruh observasi dengan rumus (Sokolova & Lapalme, 2009):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Penelitian terkini merekomendasikan penggunaan *F1-score* sebagai ukuran tambahan karena mengombinasikan *precision* dan *recall* sehingga lebih adil antar kelas (Suhaeni & Yong, 2023). *F1-score* yang tinggi menandakan model mampu mengenali kelas minoritas dengan baik. *F1-score* dihitung dengan rumus (Sokolova & Lapalme, 2009):

$$F1 = 2 \times \frac{Precission\ \times Recall}{Precission + Recall}$$

HASIL PENELITIAN DAN PEMBAHASAN

Sebelum dilakukan analisis lebih lanjut, data teks melalui tahapan pra-pemrosesan untuk memastikan kualitas dan konsistensi. Proses ini meliputi *case folding* (mengubah huruf kapital menjadi huruf kecil), penghapusan URL, penghapusan *mention* (*@user*), penggantian *hashtag* menjadi kata aslinya, pembersihan tanda baca, normalisasi spasi ganda, penghapusan *stopwords,* serta

normalisasi teks. Tahapan ini bertujuan mengubah teks menjadi lebih sederhana namun tetap mempertahankan makna utama yang diperlukan untuk analisis sentimen.

Perbedaan antara teks asli dan hasil pra-pemrosesan dapat dilihat pada Tabel 1. Tabel tersebut menampilkan contoh data sebelum dan sesudah pra-pemrosesan, sehingga terlihat jelas bagaimana komponen yang tidak relevan berhasil dihapus.

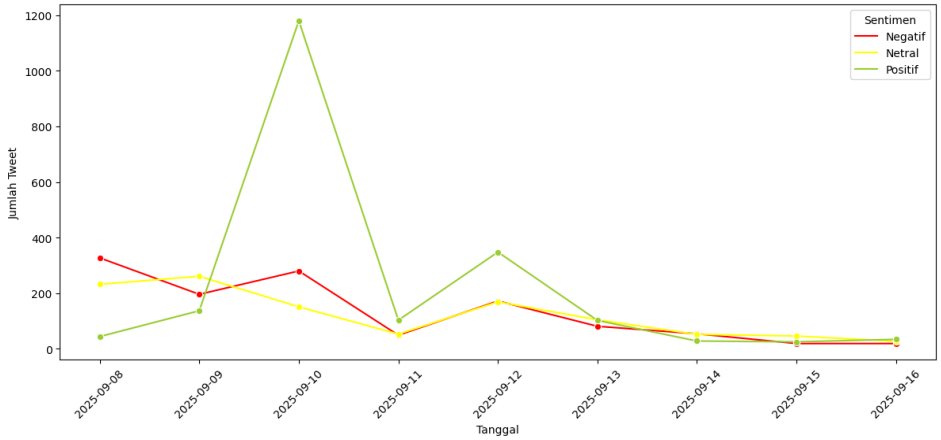
Tabel 1. Contoh Data Asli dan Hasil Pra-Pemrosesan

Data Asli	Hasil Pra-Pemrosesan
Pengamat Sebut Reshuffle Diperlukan untuk Perbaiki Kinerja Menteri Kabinet Merah Putih https://t.co/BXYG1I0I4F	pengamat sebut reshuffle diperlukan untuk perbaiki kinerja menteri kabinet merah putih
@MuhadklyAcho Semua reshuffle drama doang	reshuffle drama
Reshuffle kabinet ga di hari rabu Prabowo mematahkan tradisi Rabu reshuffle	reshuffle kabinet rabu prabowo patah tradisi rabu reshuffle
IHSG Dibuka Menguat Usai Penunjukkan Purbaya Yudhi Sadewa jadi Menkeu https://t.co/DRcV4kKQZk	ihsg buka kuat tunjuk purbaya yudhi sadewa menkeu

Wordcloud pada Gambar 1 menampilkan kata-kata yang paling sering muncul dalam percakapan publik terkait *reshuffle* Menteri Keuangan. Kata “purbaya”, “sri”, “mulyani”, dan “menkeu” mendominasi yang menunjukkan bahwa diskusi publik berfokus pada aktor utama *reshuffle*. Fenomena ini sejalan dengan fokus studi analisis sentimen dalam domain yang dikaji oleh Rodríguez-Ibánez et al. (2023), yang menunjukkan kecenderungan penelitian berpusat pada isu-isu terkait pemilu, partai politik, dan dinamika emosional dalam komunikasi kampanye.

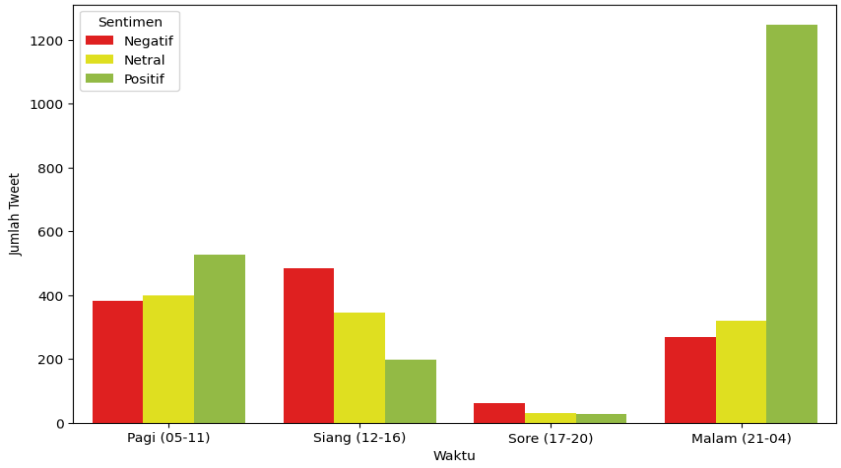
Gambar 1. *Wordcloud* Komentar Cuitan X

Tren sentimen harian pada Gambar 2 memperlihatkan lonjakan tajam sentimen positif pada 9 September 2025 yang bertepatan dengan maraknya penggunaan tagar #ThankYouSriMulyani. Pola lonjakan yang terlokalisasi pada satu hari ini mengindikasikan kemungkinan adanya orkerstrasi komunikasi, misalnya melalui kampanye terkoordinasi atau *influencer* politik. Fenomena ini sejalan dengan salah satu aplikasi penting analisis sentimen yang dikaji oleh Jim et al. (2024), yaitu mengidentifikasi dinamika komunikasi strategis, termasuk korelasi antara lonjakan sentimen dengan aktivitas akun berpengaruh dan kampanye terstruktur. Temuan ini menegaskan bahwa analisis sentimen tidak hanya mencerminkan opini organik, tetapi juga dinamika komunikasi strategis di ruang digital.



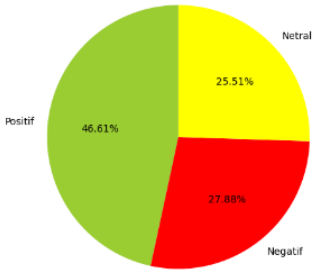
Gambar 2. Distribusi Tren Sentimen Harian

Gambar 3 menunjukkan distribusi sentimen berdasarkan waktu publikasi (pagi, siang, sore, malam). Sentimen positif mendominasi pada malam hari dengan lebih dari 1.100 cuitan, sedangkan sentimen negatif relatif tinggi pada siang hari. Sementara itu, sore hari justru memiliki intensitas cuitan paling rendah. Temuan ini mengindikasikan bahwa waktu publikasi berpengaruh pada pola ekspresi publik, dimana dukungan lebih banyak ditunjukkan pada malam hari, sedangkan kritik muncul lebih aktif pada jam produktif siang hari.



Gambar 3. Distribusi Sentimen berdasarkan Waktu

Distribusi kelas sentimen pada data awal menunjukkan kondisi yang tidak seimbang, sebagaimana terlihat pada Gambar 4. Sentimen positif mendominasi dengan proporsi 46,61%, sedangkan sentimen negatif hanya 27,88% dan sentimen netral 25,51%. Kondisi ini sejalan dengan temuan Carvalho et al. (2025) dan Alkhawaldeh et al. (2023) yang menyoroti masalah ketidakseimbangan kelas pada analisis sentimen publik. Ketidakseimbangan ini berpotensi menimbulkan bias pada model klasifikasi karena model cenderung lebih mudah memprediksi kelas mayoritas dibanding kelas minoritas.



Gambar 4. Persentase Kelas Sentimen

.

Berbeda dengan sebagian penelitian terdahulu yang mengatasi ketidakseimbangan dengan *oversampling*, *undersampling*, atau SMOTE, penelitian ini memanfaatkan LLM (GPT-4o) untuk melakukan augmentasi teks pada kelas minoritas. Strategi tersebut sejalan dengan gagasan Wei dan Zou (2019) serta Wozniak dan Kocon (2023) yang menunjukkan bahwa augmentasi berbasis LLM mampu menghasilkan teks sintesis yang lebih alami dibanding teknik manipulasi token sederhana. Dengan demikian, diharapkan model *machine learning* dapat memperoleh data pelatihan yang lebih beragam tanpa kehilangan konteks asli.

Pada tahap implementasi, augmentasi dilakukan melalui fungsi *batch_augment_with_examples* dengan pendekatan *batching* dengan ukuran 5. Setiap *batch* terdiri dari 5 teks asli yang diproses secara bersamaan untuk menghasilkan sejumlah variasi baru. Penelitian ini memberikan contoh kalimat asli dari *dataset* di dalam *prompt.* Contoh ini digunakan sebagai referensi gaya dan polaritas sentimen. Misalnya untuk sentimen negatif disertakan kalimat seperti "*presiden kita ini sebenernya bisa memimpin gak sih? reshuffle koq cuman ganti orang doang tapi kapasitasnya gak jauh sama yg sebelumnya*". Sementara itu, untuk sentimen netral diberikan contoh seperti "*ya ini valid. menurut laporan reuters dan bloomberg sri mulyani indrawati telah digantikan oleh purbaya yudhi sadewa sebagai menteri keuangan dalam reshuffle kabinet pada 8 september 2025*".

Prompt yang digunakan dalam proses augmentasi disusun secara eksplisit untuk mengarahkan model menghasilkan variasi kalimat baru tanpa menyalin instruksi atau menambahkan penomoran. Struktur *prompt* yang digunakan adalah: *"Tulis n variasi kalimat baru untuk setiap teks berikut. Jangan ulangi instruksi. Jangan sertakan penomoran atau kata tambahan. Langsung keluarkan kalimat hasilnya saja. Kalimat harus sesuai sentimen {label} (jangan ubah polaritasnya). Berikut contoh kalimat dengan sentimen {label}:... Teks yang perlu divariasikan:..."*

Pengaturan inferensi pada GPT-4o menggunakan parameter *temperature* = 0.9 yang berfungsi untuk mengontrol tingkat kreativitas atau keragaman hasil keluaran. Nilai *temperature* yang tinggi (mendekati 1) mendorong model menghasilkan kalimat yang lebih bervariasi, meskipun dengan risiko sedikit lebih

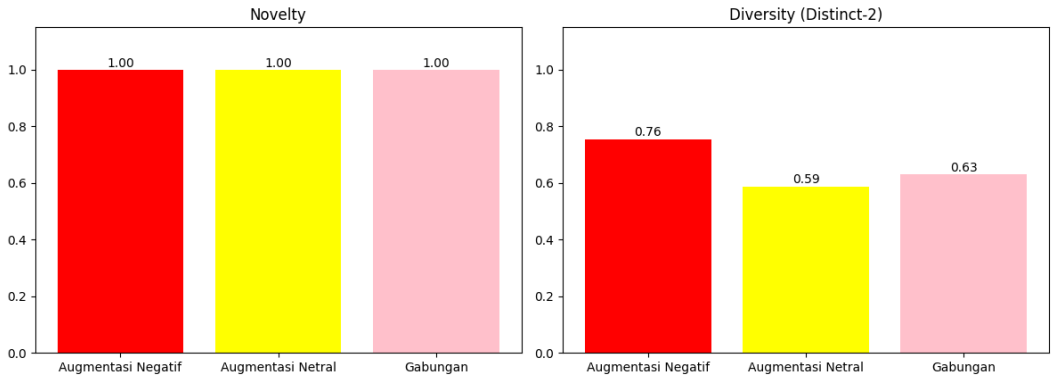
banyak deviasi dari pola asli. Nilai *temperature* yang rendah (mendekati 0) membuat keluaran lebih deterministik dan cenderung repetitif. Pemilihan nilai 0.9 pada penelitian ini bertujuan menyeimbangkan antara kebaruan (*novelty*) dan konsistensi polaritas sentimen, sehingga kalimat hasil augmentasi tetap relevan dengan kategori aslinya sekaligus cukup beragam untuk memperkaya data pelatihan.

Contoh hasil augmentasi teks ditunjukkan pada Tabel 2 yang memuat beberapa kalimat sintesis beserta label sentimennya. Tabel ini hanya menampilkan sebagian kecil dari total data hasil augmentasi yang digunakan dalam penelitian. Terlihat bahwa kalimat baru yang dihasilkan tetap konsisten dengan polaritas sentimen masing-masing.

Tabel 2. Contoh Data Sintesis dan Label Sentimen

Kalimat Sintesis	Label Sentimen
Pemilihan menteri yang baru dilantik memberikan informasi yang tidak sesuai dengan kenyataan. Baru menjabat sudah kacau, kira-kira setelah mengganti Sri Mulyani akan lebih baik ternyata sama-sama buruk.	Negatif
Penggantinya Sri Mulyani malah dapat yang nggak becus beneran.	Negatif
Menteri baru, tolong jangan teruskan kebodohan Sri Mulyani, mending kerja yang benar daripada ngasih pernyataan bodoh.	Negatif
perubahan posisi menteri keuangan dari sri mulyani indrawati ke purbaya yudi sadewa membawa dampak pada sektor keuangan. bagaimana perkembangannya sejauh ini?	Netral
Melalui berita, Sri Mulyani telah resmi mengundurkan diri dari jabatannya sebagai Menteri Keuangan dalam kabinet Merah Putih.	Netral
Menteri keuangan yang baru saja dilantik kemarin menggantikan Sri Mulyani.	Netral

Evaluasi hasil augmentasi dilakukan menggunakan dua ukuran utama yaitu *novelty* dan *diversity* seperti yang ditunjukkan pada Gambar 5. Nilai *novelty* yang diperoleh mencapai 1,00 pada semua kategori (negatif, netral, dan gabungan) yang berarti seluruh kalimat hasil augmentasi benar-benar baru dan tidak identik dengan data asli. Duffy et al. (2025) menekankan bahwa *novelty* merupakan prasyarat penting agar augmentasi benar-benar memperkaya ruang fitur dan tidak sekedar mengulang pola yang sudah ada. Temuan penelitian ini mendukung argumen tersebut dan menunjukkan bahwa GPT-4o mampu menyediakan variasi teks yang berbeda tanpa kehilangan koherensi.

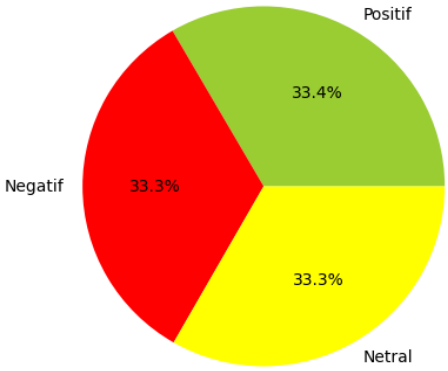


Gambar 5. Evaluasi Data Augmentasi

Nilai *diversity* diukur menggunakan *distinct*-2 yang mengindikasikan proporsi bigram (dua kata berurutan) yang unik dalam *dataset.* Hasil menunjukkan bahwa augmentasi pada kelas negatif memiliki tingkat keragaman tertinggi (0,76) diikuti oleh data gabungan (0,63), dan kelas netral (0,59). Data gabungan merujuk pada hasil penggabungan augmentasi kelas negatif dan netral sebagai keseluruhan. Perbedaan nilai *diversity* antar kelas ini dapat disebabkan oleh variasi kosakata yang lebih kaya dalam cuitan bernuansa negatif dibandingkan dengan netral.

Secara keseluruhan, evaluasi ini menegaskan bahwa proses augmentasi tidak hanya menyeimbangkan jumlah data antar kelas, tetapi juga tetap menjaga keragaman bahasa yang penting untuk meningkatkan kemampuan generalisasi model klasifikasi sentimen. Wozniak dan Kocon (2023) serta Suhaeni dan Yong (2023) menunjukkan bahwa peningkatan kualitas dan keragaman data sintesis yang dihasilkan LLM berkorelasi dengan peningkatan kemampuan generalisasi model pada tugas klasifikasi teks.

Data augmentasi untuk kelas negatif dan netral kemudian digabungkan dengan data asli sehingga menghasilkan distribusi kelas sentimen yang lebih seimbang sebagaimana terlihat pada Gambar 6. Pendekatan ini berbeda dari penelitian Setyo Nugroho et al. (2020) serta Baharuddin dan Naufal (2023) yang fokus pada penerapan *fine-tuning IndoBERT* pada tugas klasifikasi spesifik tanpa menjadikan penanganan ketidakseimbangan kelas sebagai bagian eksplisit dari metodologi mereka. Pada penelitian ini GPT-4o berperan sebagai *generator* khusus untuk kelas minoritas yang kemudian diintegrasikan ke dalam skema *fine-tuning IndoBERT*.



Gambar 6. Proporsi Kelas Sentimen di *Dataset* Gabungan

Tahap awal dalam implementasi model *IndoBERT* adalah proses tokenisasi yaitu mengubah teks mentah menjadi representasi numerik yang dapat diproses oleh model. Penelitian ini menggunakan *tokenizer* dari model *pre-trained IndoBERT-base-p1* yang tersedia pada *repository Hugging Face. Tokenizer* bertugas memecah setiap kalimat menjadi unit dasar (token), kemudian memetakan token-token tersebut ke dalam indeks numerik sesuai dengan kosakata yang dimiliki *IndoBERT*.

Pemodelan dilakukan dengan memanfaatkan arsitektur *IndoBERT* yang diimplementasikan melalui kelas *BERTSequenceClassification.* Model ini diinisialisasi dari *pretrained weights indobenchmark/indobert-base-p1* dengan jumlah label ditentukan sebanyak tiga kelas sentimen yaitu positif, negatif, dan netral. Penetapan *id2label* dan *label2id* digunakan untuk menjaga konsistensi antara representasi numerik dan label kategori sentimen dalam dataset.

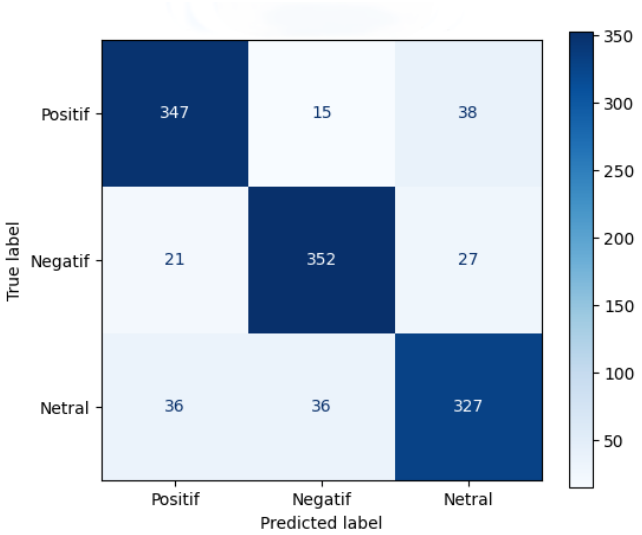
Pengaturan pelatihan model menggunakan *Hugging Face Trainer API* dengan parameter utama berupa *learning rate* sebesar 2e-5, ukuran *batch* 16 untuk data latih maupun uji, jumlah *epoch* sebanyak 3 kali iterasi penuh, serta regularisasi *weight decay* 0,01. Nilai evaluasi yang dihitung mencakup akurasi dan *F1-score* yang dirancang untuk menangkap keseimbangan performa antar kelas dalam kondisi distribusi data yang seimbang pasca augmentasi.

Tabel 3. Nilai *Training Loss* Pelatihan Model

Step	*Training Loss*
100	0,602100
200	0,437300
300	0,414500
400	0,238600
500	0,219200
600	0,211500
700	0,082900
800	0,118500
900	0,108300

Tabel 3 menunjukkan hasil pelatihan model menunjukkan bahwa nilai *training loss* menurun secara konsisten dari sekitar 0,60 pada langkah awal menjadi 0,10 di langkah terakhir (900/900 *step*). Hal ini menandakan proses optimasi berjalan baik tanpa gejala *overfitting* yang signifikan.

Confusion matrix pada Gambar 7 menunjukkan distribusi model *IndoBERT* terhadap tiga kelas sentimen yaitu positif, negatif, dan netral pada data uji. Sebagian besar data berhasil diprediksi dengan benar, misalnya 347 dari 400 cuitan positif terklasifikasi benar sebagai positif, 352 dari 400 cuitan negatif terklasifikasi benar sebagai negatif, serta 327 dari 399 cuitan netral terklasifikasi benar sebagai netral. Terdapat sejumlah kesalahan klasifikasi khususnya pada kelas netral yang relatif sering tertukar menjadi positif (36 kasus) maupun negatif (36 kasus). Hal ini mengindikasikan bahwa kalimat netral cenderung lebih sulit dibedakan karena kedekatan konteks linguistiknya dengan kelas lain.



Gambar 7. *Confusion Matrix IndoBERT*

Evaluasi model *IndoBERT* pada data uji mencapai akurasi 86% dengan nilai rata-rata *precision, recall,* dan *F1-score* yang konsisten di angka 0,86 seperti yang ditunjukkan pada Tabel 4. Kinerja ini sejalan dengan temuan Aras et al. (2024) yang juga melaporkan efektivitas *IndoBERT* dalam klasifikasi sentimen pada ulasan *e-commerce* berbahasa Indonesia.

Tabel 4. Evaluasi Klasifikasi pada Data Uji

	precision	*recall*	*f1-score*	*support*
Positif	0,86	0,87	0,86	400
Negatif	0,87	0,88	0,88	400
Netral	0,83	0,82	0,83	399
accuracy			0,86	1199
macro avg	0,86	0,86	0,86	1199
weight avg	0,86	0,86	0,86	1199

Kelas negatif menunjukkan performa terbaik dengan *F1-score* sebesar 0,88, diikuti kelas positif dengan 0,86, sedangkan kelas netral relatif lebih rendah dengan *F1-score* 0,83. Hasil ini mengindikasikan bahwa *IndoBERT* mampu melakukan klasifikasi sentimen secara andal pada teks media sosial. Temuan ini sejalan dengan Suhaeni & Yong (2023) yang menegaskan bahwa *F1-score* merupakan metrik yang lebih representatif dibanding akurasi pada *imbalanced data* karena mampu memberikan gambaran performa model yang lebih adil terhadap kelas mayoritas maupun minoritas. Dengan demikian, capaian *F1-score* yang stabil pada penelitian ini memperlihatkan bahwa pendekatan augmentasi GPT-4o berhasil membantu

menjaga keseimbangan distribusi kelas sehingga performa model tidak bias pada kelas mayoritas.

SIMPULAN

Penelitian ini berhasil menunjukkan bahwa permasalahan *imbalanced data* dalam sentimen X terkait isu *reshuffle* Menteri Keuangan dapat diatasi melalui strategi augmentasi teks menggunakan GPT-4o. Pendekatan ini mampu meningkatkan proporsi kelas menjadi seimbang sekaligus menjaga kebaruan dan keragaman data sintesis. Kondisi data yang lebih representatif menghasilkan model *fine-tuning IndoBERT* dengan performa klasifikasi yang andal dengan akurasi 86% dan nilai *F1-score* yang stabil di seluruh kelas. Temuan ini menegaskan bahwa integrasi teknik augmentasi berbasis model generatif dengan pendekatan *transformer* memberikan solusi efektif untuk mengurangi bias model akibat distribusi data yang timpang, serta memperkuat pemanfaatan analisis sentimen sebagai sarana memahami persepsi publik dalam konteks kebijakan pemerintah.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada dosen pengampu mata kuliah *Eksplorasi dan Visualisasi Data* serta Program Studi Statistika dan Sains Data IPB University yang telah memberikan dukungan, arahan, dan kesempatan dalam proses penyusunan artikel ini.

DAFTAR PUSTAKA

Alkhawaldeh, I. M., Albalkhi, I., & Naswhan, A. J. (2023). Challenges and limitations of synthetic minority oversampling techniques in machine learning. *World Journal of Methodology*, *13*(5), 373–378. https://doi.org/10.5662/wjm.v13.i5.373

Aras, S., Yusuf, M., Ruimassa, R. Y., Wambrauw, E. A. B., & Pala'langan, E. B. (2024). Sentiment Analysis on Shopee Product Reviews Using IndoBERT. *Journal of Information Systems and Informatics*, *6*(3), 1616–1627. https://doi.org/10.51519/journalisi.v6i3.814

Ashar, M. N., & Siahaan, D. O. (2024). Text Augmentation to Overcome Data Limitations in Sentiment Analysis for Bahasa Indonesia. *Proceedings of 2024 IEEE International Conference on Data and Software Engineering: Data-Driven Innovation: Transforming Industries and Societies, ICoDSE 2024*, 217–222. https://doi.org/10.1109/ICODSE63307.2024.10829895

Baharuddin, F., & Naufal, M. F. (2023). Fine-Tuning IndoBERT for Indonesian Exam Question Classification Based on Bloom's Taxonomy. *Journal of*

Information Systems Engineering and Business Intelligence, *9*(2), 253–263. https://doi.org/10.20473/jisebi.9.2.253-263

Canesta, F., & Roestam, R. (2024). Influencer pricing prognostication on social media dynamics: An advanced examination of the linear regression polynomial degree algorithm & neural networks. *MUST: Journal of Mathematics Education, Science and Technology*, *9*(2), 76–91. https://doi.org/10.30651/must.v5i1.21040

Carbonell, J. (1998). *The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries*.

Carvalho, M., Pinho, A. J., & Brás, S. (2025). Resampling approaches to handle class imbalance: a review from a data perspective. *Journal of Big Data*, *12*(1). https://doi.org/10.1186/s40537-025-01119-4

Duffy, W., O'Connell, E., McCarroll, N., Sloan, K., Curran, K., McNamee, E., Clist, A., & Brammer, A. (2025). Evaluating rule-based and generative data augmentation techniques for legal document classification. *Knowledge and Information Systems*. https://doi.org/10.1007/s10115-025-02454-x

Jim, J. R., Talukder, M. A. R., Malakar, P., Kabir, M. M., Nur, K., & Mridha, M. F. (2024). Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, *6*, 100059. https://doi.org/10.1016/j.nlp.2024.100059

Lin, C. H., & Nuha, U. (2023). Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy. *Journal of Big Data*, *10*(1). https://doi.org/10.1186/s40537-023-00782-9

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311. https://doi.org/10.3115/1073083.1073135

Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P. M. (2023). A review on sentiment analysis from social media platforms. In *Expert Systems with Applications* (Vol. 223). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2023.119862

Setyo Nugroho, K., Yullian Sukmadewa, A., Wuswilahaken, H. D., Abdurrachman Bachtiar, F., & Yudistira, N. (2021). *BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews*. https://research.google/teams/brain.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Suhaeni, C., & Yong, H. S. (2023). Mitigating Class Imbalance in Sentiment Analysis through GPT-3-Generated Synthetic Sentences. *Applied Sciences (Switzerland)*, *13*(17). https://doi.org/10.3390/app13179766

Wei, J., & Zou, K. (2019). *EDA: EASY DATA AUGMENTATION TECHNIQUES FOR BOOSTING PERFORMANCE ON TEXT CLASSIFICATION TASKS*.

Wozniak, S., & Kocon, J. (2023). From Big to Small Without Losing It All: Text Augmentation with ChatGPT for Efficient Sentiment Analysis. *IEEE International Conference on Data Mining Workshops, ICDMW*, 799–808. https://doi.org/10.1109/ICDMW60847.2023.00108

.