Jurnal Informatika: Jurnal pengembangan IT

Vol. 11, No. 2, 2026
ISSN : 2477-5126, DOI:10.30591/jpit.v11i2.10157

Generative AI vs SMOTE: Studi Kasus Penyeimbangan Data Teks pada Analisis Sentimen

Dyah Sulistyowati Rahayu [1], Iman Paryudi [2], Erin Divayaning [3], Afni Puspita Zahra [4], Arsya Yan Duribta [5]
[1,2,3,4,5]Teknik Informatika, Fakultas Teknik, Universitas Pancasila, Indonesia
[1]dyah.s.rahayu@univpancasila.ac.id, [2]iman.paryudi@univpancasila.ac.id, [3]4521210057@univpancasila.ac.id,
[4]4522210117@univpancasila.ac.id, [5]4522210117@univpancasila.ac.id,

Info Artikel

Riwayat Artikel:

Received 2026-01-11
Revised 2026-04-20
Accepted 2026-04-19

Abstract – *Imbalanced data remains a major challenge in sentiment analysis, where the dominance of positive reviews often leads to biased classification results and weak recognition of minority classes. This study aims to address the imbalance problem by applying Large Language Models (LLM) to generate synthetic negative reviews and comparing the results with the traditional SMOTE method. The research process begins with data collection through web scraping, followed by preprocessing using standard text cleaning techniques such as tokenization, stopword removal, and stemming. Augmentation is then performed with LLM to produce additional negative samples, while SMOTE is applied as a baseline method. The classification task is conducted using Support Vector Machine (SVM) with TF-IDF representation, and model performance is evaluated using accuracy, precision, recall, and F1-score. The findings show that LLM augmentation produces synthetic data highly similar to the original distribution, as confirmed by Kolmogorov-Smirnov and Wasserstein Distance tests. Furthermore, the SVM model trained with LLM-augmented data achieved higher accuracy and balanced performance compared to SMOTE, particularly in handling minority classes. In conclusion, the use of LLM provides a more effective and natural approach for text data balancing in sentiment analysis, offering significant improvement in classification quality with accuracy of 99.41%, while using SMOTE get accuracy of 98.12%.*

Keywords: *Generative AI; Imbalanced Data; LLM; Sentiment Analysis; SMOTE*

Corresponding Author:
Dyah Sulistyowati Rahayu
Email:
dyah.s.rahayu@univpancasila.ac.id

Abstrak – *Ketidakseimbangan data masih menjadi tantangan utama dalam analisis sentimen, di mana dominasi ulasan positif sering menyebabkan hasil klasifikasi bias dan kemampuan mengenali kelas minoritas menjadi lemah. Penelitian ini bertujuan untuk menerapkan Large Language Models (LLM) yang menghasilkan ulasan negatif sintetis dan membandingkannya dengan metode SMOTE. Penelitian dimulai dengan pengumpulan data melalui web scraping, dilanjutkan dengan preprocessing menggunakan teknik pembersihan teks seperti tokenisasi, penghapusan stopword, dan stemming. Augmentasi dilakukan dengan LLM untuk menambah sampel negatif, sementara SMOTE digunakan sebagai metode pembanding. Pemodelan klasifikasi dilakukan menggunakan Support Vector Machine (SVM) dengan representasi TF-IDF, dan evaluasi kinerja menggunakan metrik akurasi, precision, recall, serta F1-score. Hasil penelitian menunjukkan bahwa augmentasi dengan LLM menghasilkan data sintetis yang sangat mirip dengan distribusi asli, sebagaimana dibuktikan melalui pengujian Kolmogorov-Smirnov dan Wasserstein Distance. Selain itu, model SVM dengan data hasil augmentasi LLM mencapai akurasi lebih tinggi dan performa yang lebih seimbang dibandingkan SMOTE, terutama dalam menangani kelas minoritas. Simpulan dari penelitian ini adalah bahwa penggunaan LLM memberikan pendekatan yang lebih efektif dan alami untuk penyeimbangan data teks dalam analisis sentimen dengan akurasi 99.41% sedangkan menggunakan SMOTE memiliki akurasi 98.12%.*

Kata Kunci*: Generative AI; Data Tidak Seimbang; LLM; Analisis Sentimen; SMOTE*

I. PENDAHULUAN

Masalah data yang tidak seimbang (*imbalanced data*) masih sering muncul dalam penerapan machine learning. Hal tersebut menyebabkan model cenderung bias dan hasil prediksi menjadi kurang akurat karena jumlah data pada kelas mayoritas jauh lebih banyak dibandingkan kelas minoritas [1][2]. Salah satu metode penyeimbangan data yang populer adalah *Synthetic Minority Over-sampling Technique* (SMOTE). Metode ini bekerja dengan membuat data baru pada kelas minoritas melalui proses interpolasi antar data yang sudah ada[3][4]. Banyak penelitian menunjukkan bahwa SMOTE mampu meningkatkan performa model pada berbagai domain, seperti deteksi risiko kredit [2] [5], prediksi kesehatan [6], maupun klasifikasi asteroida berbahaya [7]. Bahkan, analisis komprehensif menunjukkan bahwa SMOTE dapat memperbaiki akurasi model secara signifikan pada dataset yang tidak seimbang [1].

Walaupun cukup populer, SMOTE memiliki beberapa kelemahan. Pertama, data sintetis yang dihasilkan kadang tidak realistis dan bisa menyebabkan masalah *overfitting* [8][9]. Kedua, metode ini kurang baik untuk data yang kompleks atau berdimensi tinggi, karena interpolasi sederhana tidak bisa menangkap pola yang lebih sulit [10][11]. Selain itu, SMOTE juga bisa menambah *noise* pada dataset, sehingga kualitas data menurun jika tidak ada langkah tambahan untuk filtering [12]. Dengan kata lain, SMOTE membantu menambah jumlah data minoritas, tetapi

kualitas data baru sering tidak sesuai dengan distribusi asli. Kelemahan mendasar dari SMOTE adalah kualitas data sintetis yang dihasilkan. Karena berbasis interpolasi sederhana, data baru sering kali tidak sepenuhnya mencerminkan distribusi asli dan berisiko menimbulkan *overfitting* atau *noise* [8][12], terutama untuk data teks, karena interpolasi antar kata atau kalimat tidak bisa menghasilkan konteks bahasa yang alami [4].

Untuk mengatasi kelemahan tersebut, banyak penelitian yang menawarkan penggunaan Generative AI. Teknologi tersebut bisa dipakai untuk membuat data sintetis yang lebih mirip dengan data asli. Generative AI dalam bentuk *Large Language Models* (LLMs) dapat digunakan untuk data teks, misalnya menghasilkan kalimat baru yang menyerupai pola bahasa pada dataset[13]. Sedangkan Generative Adversarial Networks (GANs) lebih cocok untuk data numerik, karena dapat membuat data baru yang mengikuti distribusi angka dari data asli[14][15]. Dengan cara ini, generative AI bukan hanya menambah jumlah data minoritas, tetapi juga menjaga kualitas datanya sehingga model bisa belajar lebih baik dan tidak menimbulkan permasalahan baru akibat bias data seperti pada SMOTE.

Dalam konteks penggunaan data teks, *Large Language Models* (LLM) menawarkan pendekatan yang lebih sesuai. LLM mampu menghasilkan kalimat baru yang menyerupai pola bahasa asli, sehingga data sintetis yang dihasilkan lebih alami dan relevan. Penelitian terbaru menunjukkan bahwa LLM dapat digunakan untuk membuat sampel tekstual yang realistis dalam konteks dataset tidak seimbang, dan hasilnya lebih baik dibandingkan teknik oversampling tradisional [13].

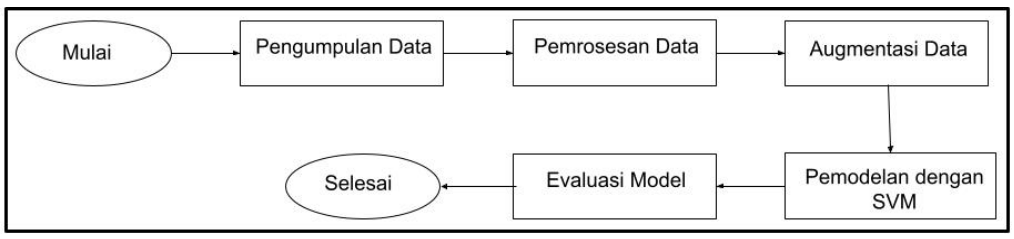
Meskipun penelitian yang memanfaatkan LLM sebagai penyeimbang data sudah ada, namun penggunaannya pada analisis sentimen secara langsung belum banyak dibahas. Oleh karena itu, berdasarkan penelitian sebelumnya, penelitian ini secara khusus menawarkan penggunaan LLM untuk mengatasi *imbalance data* pada data teks dalam analisis sentimen dengan kasus riil. Dengan memanfaatkan kemampuan LLM, data minoritas pada kategori sentimen tertentu dapat diperbanyak dengan cara yang lebih alami, sehingga model analisis sentimen bisa bekerja lebih baik. Hal ini diharapkan dapat meningkatkan akurasi dan mengurangi bias pada hasil klasifikasi.

Kebaruan dari penelitian ini terletak pada penggunaan *Large Language Models* (LLM) untuk mengatasi masalah imbalance data pada data teks dalam analisis sentimen. Pendekatan ini menawarkan cara baru dalam menghasilkan data sintetis yang lebih alami dan sesuai dengan pola bahasa yang ada pada dataset teks. Berbeda dengan penelitian sebelumnya yang menggunakan SMOTE untuk menyeimbangkan data teks sentimen [4], penelitian ini tidak hanya menambah jumlah data minoritas tetapi juga menjaga kualitas bahasa yang dihasilkan. Dengan memanfaatkan kemampuan LLM, data minoritas pada kategori sentimen tertentu dapat diperbanyak tanpa harus bergantung pada interpolasi sederhana seperti pada SMOTE[16]. Oleh karena itu, penelitian ini memberikan kontribusi baru dalam bidang penyeimbangan data teks dengan menawarkan pendekatan berbasis generative AI-LLM yang lebih sesuai untuk konteks analisis sentimen. Kebaruan yang lainnya yaitu terletak pada penggunaannya yang langsung pada klasifikasi sentimen, sehingga pendekatan yang diusulkan dapat langsung diukur performanya dibandingkan dengan penerapan pada model lainnya dengan kasus yang sama[17]. Tingkat kemiripan data sintetis terhadap data asli juga diukur.

Penelitian ini bertujuan untuk mengaplikasikan LLM untuk menghasilkan data sintetis pada kelas minoritas dalam dataset teks. Selain itu, penelitian ini juga mengukur hasil LLM dari tingkat kemiripan data asli terhadap data sintetisnya. Terakhir, penelitian ini membandingkan akurasi hasil pemodelan yang menggunakan data hasil augmentasi menggunakan LLM dan yang menggunakan SMOTE.

II. METODE

Tahapan penelitian pada penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. Tahapan penelitian.

A. Pengumpulan Data

Web scraping dari situs *Shopee.co.id* untuk pengumpulan data penelitian dilakukan pada halaman produk *Wardah Glasting Liquid Lip*. Tautan produknya adalah https://shopee.co.id/NEW!-WARDAH-Glasting-Liquid-Lip-Hi-Pigmented-Glass-Color-Ringan-Tidak-Lengket-Transferproof-Makeup.

B. Pemrosesan Data

Tahap pemrosesan data awal ini untuk menyiapkan data ulasan agar dapat dianalisis dengan baik. Proses ini dilakukan untuk membersihkan, menormalkan, serta mengubah teks mentah menjadi format yang lebih terstruktur. Dalam praktiknya digunakan beberapa library Python yang umum dipakai untuk pengolahan teks bahasa Indonesia, seperti *sastrawi*, *nltk*, *pandas*, *numpy*, *re*, dan *transformers*.

Data hasil *scraping* mula-mula disimpan dalam format CSV, lalu dimuat ke dalam objek DataFrame. Untuk memastikan data terbaca dengan benar, beberapa baris awal ditampilkan menggunakan fungsi df.head(). Selanjutnya, fungsi df.info() dipakai untuk melihat jumlah entri, nama kolom, tipe data, serta apakah ada nilai kosong. Fungsi df.describe() digunakan untuk menampilkan ringkasan statistik, misalnya nilai minimum, maksimum, rata-rata, dan kuartil pada kolom numerik seperti rating.

Tahap pembersihan dilakukan dengan dua cara. Pertama, df.drop_duplicates() digunakan untuk menghapus baris yang sama persis agar setiap ulasan tetap unik. Kedua, df.dropna() dipakai untuk menghapus baris yang memiliki nilai kosong sehingga data yang dianalisis lengkap.

Proses berikutnya meliputi: mengubah teks ke huruf kecil (*lowercase*), melakukan tokenisasi untuk memecah kalimat menjadi kata, menghapus karakter non-teks, menghilangkan *stopwords* yang tidak berpengaruh pada analisis sentimen, serta melakukan *stemming* dengan library Sastrawi agar kata dikembalikan ke bentuk dasar. Tabel 1 menunjukkan data sebelum dan setelah pemrosesan.

TABEL 1
DATA SEBELUM DAN SESUDAH PEMROSESAN

Data sebelum pemrosesan

	username	date	comment	rating
0	_uchaaa	2024-01-16 14:38:13	Performa:9/10\nTekstur:creamy, glossy\nCocok U...	5
1	d*****9	2024-01-28 10:14:02	Efek:surprisingly gak bikin bibir kering\nKema...	5
2	_uchaaa	2024-01-16 14:41:44	Tekstur:creamy glossy\nPerforma:9.5/10\nCocok ...	5
3	syanachan	2024-01-12 14:00:53	Cocok Untuk:cool undertone\nTekstur:lip cream ...	5
4	bellanandyaf	2024-02-06 06:55:56	Cocok Untuk:semua tone kulit\nPerforma:longlas...	5
...	...	...	...	...
18295	poopuri__	2024-01-26 14:01:47	BAGUS! Ibh bagus dr merk sblh ya hehe.. soalny...	5
18296	w*****f	2024-09-15 12:26:04	Tekstur:matte\nWarna:bagus\nCocok untuk:semua ...	4
18297	alestari09	2024-02-08 15:21:50	Tekstur:tidak terlalu cair\nPerforma:bagus\nCo...	5
18298	d*****2	2024-12-16 05:24:15	Tekstur:baguss\nWarna:baguss\nCocok untuk:semu...	5
18299	lyd_rn	2024-02-14 21:23:09	Efek:bagus\nManfaat:mempercantik bibir\nPengal...	5

18300 rows × 4 columns

Data setelah pemrosesan

df

	username	date	rating	review
0	_uchaaa	2024-01-16 14:38:13	5	performa teksturcreamy glossy cocok untukbibir...
1	d*****9	2024-01-28 10:14:02	5	efeksurprisingly gak bikin bibir kering kemasa...
2	_uchaaa	2024-01-16 14:41:44	5	teksturcreamy glossy performa cocok untukbisa ...
3	syanachan	2024-01-12 14:00:53	5	cocok untukcool undertone teksturlip cream oil...
4	bellanandyaf	2024-02-06 06:55:56	5	cocok untuksemua tone kulit performalonglastin...
...	...	...	...	...
9145	u*****2	2024-10-03 05:31:51	4	warnanyaa merah ternyataa ngestain lengket git...
9146	imurday	2024-12-30 23:46:18	5	warna bagus tahan ringan ga berat bibir
9147	m*****7	2024-04-16 12:05:33	5	teksturcair kental cocok untukorang dewasa war...
9148	v*****i	2024-11-09 12:18:53	4	teksturagak lengket warnalebih gelap ekspektas...
9149	k*****6	2024-08-11 13:45:06	5	sukakk rekomendid glowsy pakai bibir cantik ba...

3097 rows × 4 columns

C. Augmentasi Data dengan LLM

Dataset yang digunakan memiliki ketidakseimbangan, dimana jumlah ulasan positif jauh lebih banyak dibandingkan ulasan negatif. Oleh karena itu, tahap ini berfokus pada penyeimbangan kelas dengan augmentasi berbasis LLM. Model LLM digunakan untuk menghasilkan ulasan negatif sintetis yang menyerupai data asli, baik dari sisi gaya bahasa maupun konteks produk. Data hasil augmentasi kemudian digabungkan dengan dataset utama sehingga distribusi kelas menjadi lebih seimbang.

Secara teknis, proses augmentasi dilakukan dengan memanfaatkan LLM yang telah dilatih secara umum pada data teks dalam jumlah besar. Model yang digunakan adalah GPT-3.5, karena memiliki kemampuan memahami konteks dan menghasilkan teks yang koheren serta menyerupai ulasan asli. Untuk menghasilkan data sintetis, digunakan prompt engineering berupa instruksi eksplisit yang meminta model menuliskan ulasan negatif mengenai produk Wardah Glasting Liquid Lip di platform Shopee. Prompt dirancang agar model tetap menjaga relevansi dengan produk, menggunakan gaya bahasa yang natural, dan menekankan aspek negatif seperti ketahanan, tekstur, atau pengalaman penggunaan.

Mekanisme generasi dilakukan dengan pengaturan parameter:

a. *Temperature* = 0.7, untuk menjaga keseimbangan antara kreativitas dan konsistensi sehingga ulasan tidak terlalu acak tetapi tetap bervariasi.
b. Top-p (*nucleus sampling*) = 0.9, agar model memilih kata-kata dari distribusi probabilitas yang lebih terfokus namun tetap memungkinkan variasi.
c. *Max tokens* = 100–150, menyesuaikan panjang ulasan agar serupa dengan data asli di Shopee.
d. *Number of outputs* per prompt = 5, sehingga dari satu instruksi dapat dihasilkan beberapa variasi ulasan negatif.

Dengan kombinasi model, prompt, dan parameter tersebut, dihasilkan kumpulan ulasan sintetis yang kemudian melalui tahap verifikasi manual untuk memastikan kesesuaian konteks dan kualitas bahasa. Setelah lolos verifikasi, ulasan sintetis ditambahkan ke dataset utama sehingga distribusi kelas positif dan negatif menjadi lebih seimbang, yang pada akhirnya meningkatkan performa model klasifikasi sentimen.

D. Pemodelan dengan SVM

Data yang sudah diproses kemudian digunakan untuk pemodelan dengan algoritma Support Vector Machine (SVM). Analisis sentimen dengan model SVM digunakan karena memiliki performa yang baik dalam penelitian sebelumnya yang sejenis[10]. Metode Term Frequency-Inverse Document Frequency (TF-IDF) digunakan untuk memperoleh representasi teks agar model dapat membedakan sentimen dengan lebih tepat.

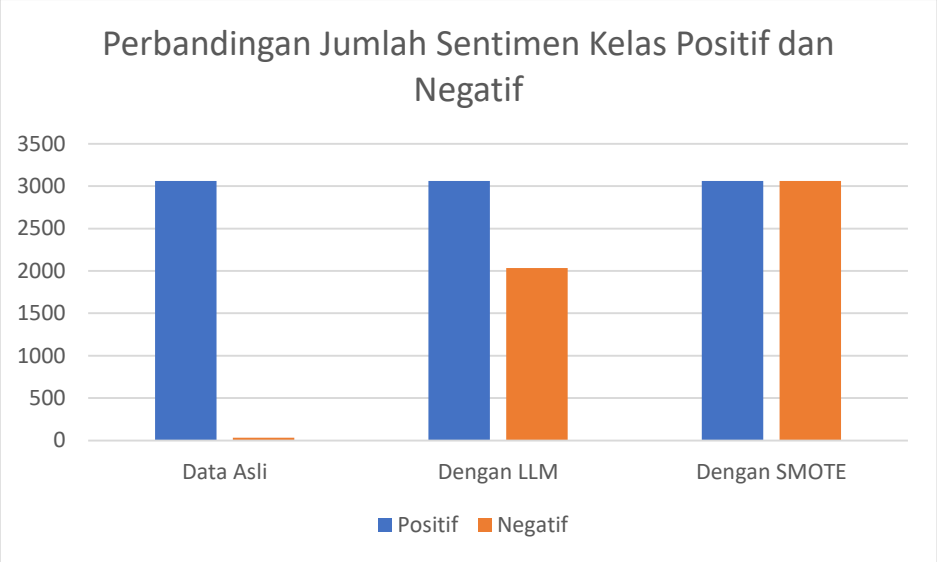
E. Evaluasi Model

Model yang dihasilkan dievaluasi menggunakan teknik cross-validation. Pada tahap ini, dilakukan tiga skema pembagian data latih dan data uji yaitu dengan rasio yang umum digunakan pada penelitian terdahulu [18]. Rasio yang digunakan yaitu 80:20, 70:30, dan 60:40. Metrik evaluasi yang digunakan meliputi akurasi, precision, recall, dan F1-score untuk menilai kinerja klasifikasi secara menyeluruh.

III. HASIL DAN PEMBAHASAN

A. Data

Dari hasil pengumpulan data, terdapat 3.097 ulasan dengan ulasan positif mendominasi sebanyak 3.062, sedangkan ulasan negatif hanya berjumlah 35. Ketidakseimbangan ini menunjukkan kondisi data yang tidak seimbang dan berpotensi menimbulkan bias pada model klasifikasi. Model yang dilatih dengan data seperti ini biasanya lebih mudah mengenali pola dari kelas mayoritas, tetapi kesulitan dalam mendeteksi kelas minoritas.



Gambar 2. Perbandingan jumlah sentimen

Untuk mengatasi masalah tersebut, penelitian ini menggunakan LLM untuk membuat 2.000 ulasan negatif sintetis. Jumlah ini dipilih secara hati-hati agar proporsi data tetap realistis dan tidak membuat dataset menjadi terlalu besar atau tidak seimbang secara berlebihan [19][20]. Selain itu, pembatasan jumlah data sintetis juga bertujuan mengurangi risiko *overfitting*. Kondisi ini menyebabkan model kehilangan kemampuan generalisasi terhadap data baru. Pemilihan jumlah data sintetis ini sejalan dengan prinsip pendekatan *partial oversampling*, yaitu bukan menyamakan jumlah kelas sepenuhnya agar distribusi alami tetap terjaga dan proses pelatihan tetap efisien. Studi sebelumnya menunjukkan bahwa penambahan data sintetis dari LLM secara berlebihan justru dapat menyebabkan masalah seperti *model collapse*, bias, dan penurunan kemampuan generalisasi [19][20]. Oleh karena itu, kontrol terhadap volume augmentasi menjadi hal yang penting. Proses augmentasi menghasilkan jumlah total data ulasan menjadi 5.097 dari sebelumnya berjumlah 3.097 dengan distribusi kelas yang lebih seimbang. Untuk keperluan pembanding, digunakan SMOTE untuk membuat kelas seimbang menjadi 3062 data untuk masing-masing kelas sehingga proporsinya masing-masing 50%. Jumlah kelas akhir ditunjukkan pada Gambar 2. Dataset hasil augmentasi dan juga SMOTE ini kemudian digunakan dalam pelatihan model klasifikasi untuk analisis sentimen.

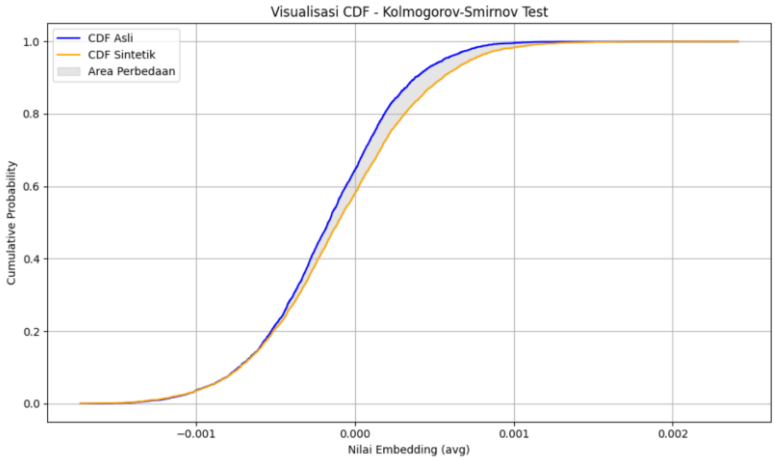
B. Pengujian Kemiripan Data Sintetis dari Data Asli

Pengujian kesamaan antara data sintetis dan data asli dilakukan dengan dua cara, yaitu Kolmogorov-Smirnov (KS) dan Wasserstein Distance (WD). Proses perhitungan metode KS yaitu dengan mengurutkan kedua sampel lalu menghitung fungsi distribusi kumulatif (CDF) masing-masing, serta dengan menghitung selisih terbesar antara kedua CDF. Jika selisih ini besar, berarti ada perbedaan yang signifikan. Sedangkan pada metode WD, kedua vektor yang dibandingkan harus memiliki panjang yang sama. Proses perhitungan Wasserstein Distance dilakukan dengan random sampling sesuai panjang minimum, kemudian mengurutkan vektor, dan menghitung rata-rata selisih absolut antar elemen yang bersesuaian. Nilai tersebut menjadi Wasserstein Distance, yang menunjukkan seberapa jauh distribusi data sintetis berbeda dari data asli.

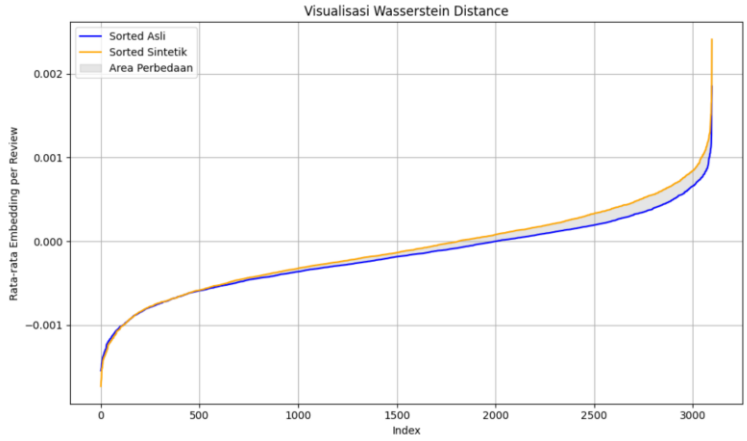
Pengujian KS pada distribusi rata-rata embedding menunjukkan nilai KS Statistic sebesar 0,0802 terhadap data sintetis. Nilai ini menandakan adanya perbedaan distribusi namun cukup kecil dan tidak signifikan. Secara umum, KS Statistic berada pada rentang 0 sampai 1, yang mana nilai semakin mendekati 0 berarti tingkat kemiripan distribusinya tinggi, sedangkan nilai di atas 0,2 biasanya menunjukkan perbedaan yang cukup jelas. Karena itu, nilai 0,0802 masih dianggap wajar dan memperlihatkan bahwa data sintetis cukup mirip dengan data asli.

Selain itu, hasil Wasserstein Distance yang diperoleh adalah 0,0001, yang berarti bahwa rata-rata perbedaan antara kedua distribusi sangat kecil. Wasserstein Distance (Earth Mover's Distance), mengukur rata-rata perbedaan distribusi. Nilai kurang dari 0,01 menunjukkan perbedaan yang kecil, sedangkan nilai di atas 0,1 menandakan kemiripan yang sangat kecil. Melalui pengukuran ini dengan hasil 0,0001 dapat disimpulkan bahwa data sintetis memiliki sifat yang sangat mirip dengan data asli berdasarkan representasi vektor kalimat.

Gambar 3 dan Gambar 4 masing-masing menunjukkan visualisasi hasil tes menggunakan metode Kolmogorov-Smirnov dan Wasserstein Distance. Dari kedua gambar tersebut terlihat bahwa kurvanya hampir sama. Hal tersebut menunjukkan kemiripan antara data sintetis yang dihasilkan oleh augmentasi menggunakan LLM terhadap data aslinya.



Gambar 3. Visualisasi hasil tes menggunakan Kolmogorov–Smirnov



Gambar 4. Visualisasi hasil tes menggunakan Wasserstein Distance

C. Evaluasi Model

Pada bagian ini dilakukan proses pemodelan terhadap data yang telah ditambah menggunakan LLM. Sebagai pembanding, pemodelan juga dijalankan pada data yang diperluas dengan metode SMOTE. Kedua pendekatan diuji menggunakan algoritma SVM untuk melihat perbedaan kinerjanya.
Secara teknis, konfigurasi Support Vector Machine (SVM) yang digunakan dalam penelitian ini adalah sebagai berikut:

a. Kernel: digunakan kernel linear, karena sesuai dengan karakteristik data teks yang direpresentasikan dengan TF-IDF. Kernel linear memungkinkan pemisahan kelas positif dan negatif secara efisien tanpa transformasi non-linear yang kompleks.
b. Parameter C (regularisasi): nilai C = 1.0 dipilih untuk menyeimbangkan antara margin yang lebar dan kesalahan klasifikasi. Nilai ini menjaga agar model tidak terlalu ketat (overfitting) maupun terlalu longgar (underfitting).
c. Tuning parameter: dilakukan eksperimen dengan beberapa nilai C (misalnya 0.1, 1, dan 10) untuk melihat pengaruhnya terhadap akurasi, namun hasil terbaik diperoleh pada C = 1.0.

Dengan konfigurasi tersebut, SVM mampu memberikan hasil yang stabil dalam klasifikasi sentimen, baik pada dataset hasil augmentasi LLM maupun dataset hasil SMOTE. Hasil pemodelan dengan data hasil augmentasi LLM menunjukkan akurasi sebesar 99,41% yang menunjukkan kemampuan klasifikasi yang baik. Gambar 5 menunjukkan bahwa pada F1-score kelas negatif mencapai 0.9 dengan precision 1.00 dan recall 0.99. Sedangkan pada kelas positif, F1-score 1 dengan precision sebesar 0.99 dan recall 1.00. Rata-rata tertimbang dari seluruh metrik juga berada pada nilai yang sangat baik, memperlihatkan bahwa model dapat mengenali kedua kelas secara seimbang dan akurat.

Laporan Klasifikasi:

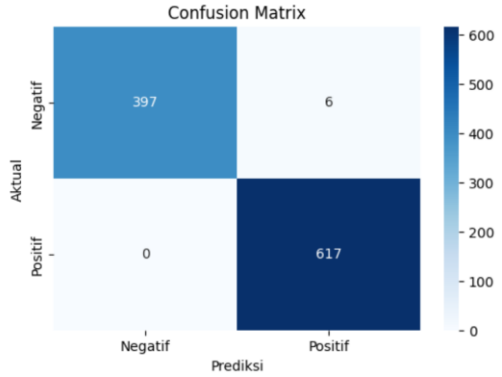
	precision	recall	f1-score	support
negatif	1.00	0.99	0.99	403
positif	0.99	1.00	1.00	617
accuracy			0.99	1020
macro avg	1.00	0.99	0.99	1020
weighted avg	0.99	0.99	0.99	1020

Gambar 5. Hasil pemodelan SVM dengan LLM

Sebagai bagian dari evaluasi, dibuat confusion matrix untuk melihat jumlah prediksi benar maupun salah pada tiap kelas. Dari hasil yang ditampilkan pada Gambar 6, diperoleh 617 True Positive (TP), 397 True Negative (TN), 6 *False Positive* (FP), dan 0 *False Negative* (FN). Temuan ini menunjukkan bahwa model mampu mengenali kelas positif dengan sangat baik, serta hanya menghasilkan kesalahan kecil pada kelas negatif.

Setelah dilakukan penerapan SMOTE pada data uji, model klasifikasi yang dibangun dengan algoritma SVM memperoleh tingkat akurasi sebesar 98,12% (0.9812). Angka ini menunjukkan bahwa performa model tergolong sangat baik. Untuk kelas negatif, nilai precision tercatat 0.97 dengan recall 0.99 sehingga menghasilkan F1-score 0.98. Sedangkan pada kelas positif, precision mencapai 0.99 dan recall 0.97 dengan F1-score 0.98 (lihat Gambar 7). Secara keseluruhan, nilai rata-rata tertimbang dari metrik evaluasi juga tinggi, menandakan bahwa model mampu bekerja cukup seimbang dalam mengenali kedua kelas.

Confusion matrix pada Gambar 8 memberikan hasil prediksi 610 True Positive (TP), 592 True Negative (TN), 6 False Positive (FP), dan 17 False Negative (FN). Secara umum, hasil ini memperlihatkan bahwa model memiliki kinerja yang solid dengan tingkat kesalahan yang relatif kecil. Meski demikian, masih ada kekeliruan pada sebagian data positif yang belum dapat diidentifikasi dengan sempurna. Dari kedua pemodelan tersebut, didapatkan perbandingan kinerja LLM dan SMOTE seperti yang ditunjukkan pada Tabel 2.



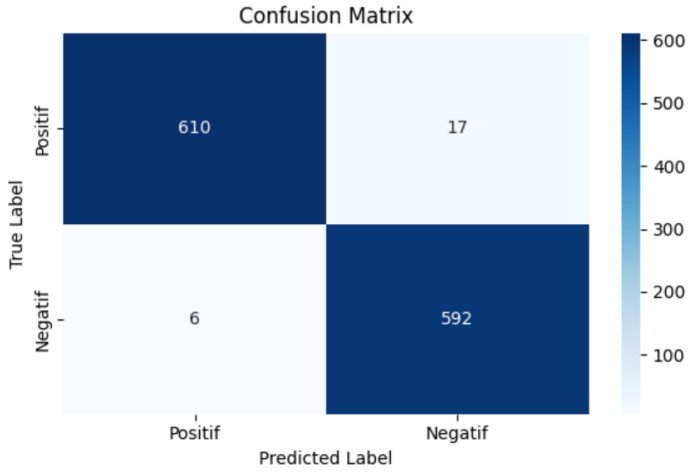
Gambar 6. *Confusion Matrix* dari hasil pemodelan SVM dengan LLM.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.99      0.98       598
           1       0.99      0.97      0.98       627

    accuracy                           0.98      1225
   macro avg       0.98      0.98      0.98      1225
weighted avg       0.98      0.98      0.98      1225
```

Gambar 7. Hasil pemodelan SVM dengan SMOTE.



Gambar 8. *Confusion Matrix* dari hasil pemodelan SVM dengan SMOTE.

TABEL 2
PERBANDINGAN HASIL EVALUASI SVM DENGAN LLM DAN SMOTE

No.	Metrik	SVM + LLM	SVM + SMOTE
1.	*Accuracy*	99,41% (0.994)	98,12% (0.9812)
2.	*Precision*	0.99 / 1.00	0.99 / 0.97
3.	*Recall*	1.00 / 0.99	0.97 / 0.99
4.	*F1-score*	1.00 / 0.99	0.98 / 0.98

D. Pembahasan

Hasil evaluasi model tersebut juga dapat dibandingkan dengan penelitian terdahulu untuk menunjukkan performa model SVM+LLM yang diuji. Penelitian Oktariansyah (2024) menggunakan IndoBERT dengan Synonym Replacement dan Back Translation memiliki akurasi tertinggi 82% dalam penanganan imbalance data[21]. Penggunaan IndoBert dengan XLM-RoBERTa memiliki akurasi 92%[22]. Algoritma BiLSTM digunakan oleh Muzakkir (2023) yang menggabungkannya dengan word embedding GloVe. Penelitian ini mendapatkan akurasi maksimal 97.5% [23]. Meskipun perbandingan tersebut memiliki data yang berbeda dengan pengujian penelitian ini, namun dengan melihat karakter data yang tidak seimbang dan penggunaan model *machine learning* yang sama, penggunaan LLM dalam analisis sentimen dapat dijadikan rujukan bagi penelitian lain yang berjalan.

Dengan kemampuan memahami konteks semantik, LLM tidak hanya menambah jumlah data minoritas tetapi juga menjaga kualitas bahasa, sehingga model analisis sentimen dapat belajar lebih seimbang. Kemiripan data sintetis terhadap data asli sangat tinggi yaitu Wasserstein Distance sebesar 0,0001.

Meskipun secara angka akurasi LLM dengan model SVM lebih baik, namun ada bias penilaian akibat jumlah data yang berbeda antara yang diujikan pada SMOTE dengan LLM. Perbedaan jumlah data kelas pada LLM menunjukkan keunggulan hasil datanya meskipun rasio kelasnya tidak persis 50:50. Perbandingan dengan penelitian lain sebagai baseline juga dapat memberikan bias terhadap hasil karena data yang digunakan berbeda.

Selain itu, pendekatan generatif lain seperti Generative Adversarial Networks (GANs) juga terbukti efektif dalam menghasilkan data sintetis yang lebih realistis pada domain numerik, misalnya lalu lintas [14], keuangan [15], dan kesehatan [24][25]. Hal ini memperkuat argumen bahwa generative AI, baik LLM untuk teks maupun GANs untuk data numerik, mampu mengatasi keterbatasan SMOTE dengan cara yang lebih fleksibel dan berkualitas tinggi.

Dengan demikian, alasan kuat mengapa LLM bisa lebih baik daripada SMOTE adalah karena LLM mampu menghasilkan data sintetis yang tidak hanya menambah jumlah kelas minoritas, tetapi juga menjaga kualitas semantik dan konteks bahasa. Hal ini sangat penting dalam analisis sentimen, di mana makna kata dan kalimat tidak bisa direpresentasikan dengan interpolasi sederhana seperti pada SMOTE.

IV. SIMPULAN

Penelitian ini bertujuan untuk mengatasi masalah ketidakseimbangan data pada analisis sentimen dengan memanfaatkan *Large Language Models* (LLM) sebagai alternatif dari metode SMOTE. Berdasarkan hasil yang diperoleh, augmentasi data menggunakan LLM mampu menghasilkan data sintetis yang memiliki distribusi sangat mirip dengan data asli, sebagaimana ditunjukkan oleh nilai Kolmogorov-Smirnov dan Wasserstein Distance yang rendah. Pemodelan dengan algoritma SVM menunjukkan bahwa data hasil augmentasi LLM memberikan akurasi dan keseimbangan klasifikasi yang lebih tinggi dibandingkan dengan SMOTE, yaitu 99.41% dibandingkan 98.12%. Temuan ini menegaskan bahwa LLM dapat menjadi pendekatan yang lebih efektif untuk penyeimbangan data teks dalam analisis sentimen. Ke depan, penelitian lanjutan dapat diarahkan pada eksplorasi penggunaan LLM dalam domain teks lain yang lebih kompleks, serta pengembangan strategi integrasi dengan metode generatif lain seperti GANs untuk memperluas penerapan pada data numerik maupun multimodal.

UCAPAN TERIMAKASIH

Penulis berterima kasih kepada Fakultas Teknik Universitas Pancasila atas dukungan hibah penelitian yang diberikan melalui program Hibah Kelompok Riset Fundamental (KRF) No.1384/D/FTUP/VIII/2025.

DAFTAR PUSTAKA

[1] B. Santhosh Kumar, P. Praveen Yadav, and P. Penchala Prasad, "Performance Analysis of Machine Learning Algorithms on Imbalanced Datasets Using SMOTE Technique," in *Proceedings of the 5th International Conference on Data Science, Machine Learning and Applications. Lecture Notes in Electrical Engineering.*, Springer, 2025, pp. 147–156. doi: 10.1007/978-981-97-8031-0_15.

[2] N. Matar, B. Sowan, and A. Al-Jaber, "Evaluating Models Performance for Credit Risk Detection for Imbalanced Data," in *2024 2nd International Conference on Cyber Resilience (ICCR)*, IEEE, Feb. 2024, pp. 1–6. doi: 10.1109/ICCR61006.2024.10532912.

[3] S. Călin, "Handling Imbalanced Data: The SMOTE Technique," in *2025 17th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, IEEE, Jun. 2025, pp. 1–5. doi: 10.1109/ECAI65401.2025.11095450.

[4] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, "Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks," *Applied Sciences*, vol. 13, no. 6, p. 4006, Mar. 2023, doi: 10.3390/app13064006.

[5] N. N. A. Nanda, Y. Farida, and W. D. Utami, "Implementation of SMOTE to Improve the Performance of Random Forest Classification in Credit Risk Assessment in Banking," *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 9, no. 2, pp. 158–177, Jul. 2025, doi: 10.29407/intensif.v9i2.23930.

[6] C. Decaro, G. B. Montanari, M. Bianconi, and G. Bellanca, "Prediction of hematocrit through imbalanced dataset of blood spectra," *Healthc. Technol. Lett.*, vol. 8, no. 2, pp. 37–44, Apr. 2021, doi: 10.1049/htl2.12006.

[7] M. S. Kennanya, T. Meena, M. S. Pravardhitha, and A. S. Vignesh, "Classification of Potentially Hazardous Asteroids Using Artificial Neural Networks and Over Sampling Techniques," in *2023 Global Conference on Information Technologies and Communications (GCITC)*, IEEE, Dec. 2023, pp. 1–6. doi: 10.1109/GCITC60406.2023.10426106.

[8] A. H. Butt, Z. Khan, A. Khan, H. Ghazanfar, R. Zgheib, and F. Kamalov, "Performance of Sampling Methods on Imbalanced Data: Comparative Analysis," in *2024 Advances in Science and Engineering Technology International Conferences (ASET)*, IEEE, Jun. 2024, pp. 1–6. doi: 10.1109/ASET60340.2024.10708760.

[9] M. M. K. Dandu, J. Jain, S. Vijayabaskar, P. Goel, A. Shivarudra, and S. Bhatt, "Assessing the Impact of Data Imbalance on the Predictive Performance of Machine Learning Models," in *2024 7th International Conference on Contemporary Computing and Informatics (IC3I)*, IEEE, Sep. 2024, pp. 1062–1068. doi: 10.1109/IC3I61595.2024.10829313.

[10] H. Cheng, "Support Vector Machine with SMOTE Based on Correlated Covariates," in *2023 2nd International Conference on Automation, Robotics and Computer Engineering (ICARCE)*, IEEE, Dec. 2023, pp. 1–4. doi: 10.1109/ICARCE59252.2024.10492566.

[11] C. Zhang, J. Song, Z. Pei, and J. Jiang, "An Imbalanced Data Classification Algorithm of De-noising Auto-Encoder Neural Network Based on SMOTE," *MATEC Web of Conferences*, vol. 56, p. 01014, Apr. 2016, doi: 10.1051/matecconf/20165601014.

[12] Z. Wei and Y. Chen, "NLKF-SMOTE: A Novel Noise-Filtering SMOTE Without Nearest Neighbor Parameter K for Oversampling," in *Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence*, New York, NY, USA: ACM, Dec. 2024, pp. 153–159. doi: 10.1145/3709026.3709035.

[13] S. Gopali, F. Abri, A. Siami Namin, and K. S. Jones, "The Applicability of LLMs in Generating Textual Samples for Analysis of Imbalanced Datasets," *IEEE Access*, vol. 12, pp. 136451–136465, 2024, doi: 10.1109/ACCESS.2024.3463400.

[14] M.-Y. Chen, H.-S. Chiang, and W.-K. Huang, "Efficient Generative Adversarial Networks for Imbalanced Traffic Collision Datasets," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 19864–19873, Oct. 2022, doi: 10.1109/TITS.2022.3162395.

[15] M. Jiang, Y. Liang, S. Han, K. Ma, Y. Chen, and Z. Xu, "Leveraging Generative Adversarial Networks for Addressing Data Imbalance in Financial Market Supervision," in *Proceedings of the 2024 5th International Conference on Big Data Economy and Information Management*, New York, NY, USA: ACM, Dec. 2024, pp. 651–656. doi: 10.1145/3724154.3724263.

[16] C. Paramita, C. S. Simbolon, A. S. Pamungkas, J. M. Triono, E. P. Widi Utomo, and E. R. Subhiyakto, "Analisis Pengaruh SMOTE terhadap Kinerja Model KNN untuk Prediksi Risiko Stroke," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 10, no. 4, pp. 978–988, Sep. 2025, doi: 10.30591/jpit.v10i4.8809.

[17] F. Mahardika, "Analisis Sentimen pada Implementasi Pembelajaran Berbasis AI : Studi Kasus Persepsi Mahasiswa dan Dosen di Institusi Swasta," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 11, no. 1, pp. 197–204, Jan. 2026, doi: 10.30591/jpit.v11i1.9069.

[18] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, "Trade-off between training and testing ratio in machine learning for medical image processing," *PeerJ Comput. Sci.*, vol. 10, 2024, doi: 10.7717/PEERJ-CS.2245.

[19] Jie Chen, Yupeng Zhang, Bingning Wang, Wayne Xin Zhao, Ji-Rong Wen, and Weipeng Chen, "Unveiling the Flaws: Exploring Imperfections in Synthetic Data and Mitigation Strategies for *Large Language Models*," *Arxiv Computation and Language*, 2024.

[20] Kareem Amin, Sara Babakniya, Alex Bie, Weiwei Kong, Umar Syed, and Sergei Vassilvitskii, "Escaping Collapse: The Strength of Weak Data for Large Language Model Training," *Arxiv Machine Learning*, 2025.

[21] I. A. Oktariansyah, F. R. Umbara, and F. Kasyidi, "Klasifikasi Sentimen Untuk Mengetahui Kecenderungan Politik Pengguna X Pada Calon Presiden Indonesia 2024 Menggunakan Metode IndoBert," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 2, pp. 636–648, Sep. 2024, doi: 10.47065/bits.v6i2.5435.

[22] A. Syarifuddin, "Deep Learning-Based Sentiment Analysis of Islamic Boarding School Google Reviews Using IndoBERT Variants and XLM-RoBERTa," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 11, no. 1, pp. 152–161, Jan. 2026, doi: 10.30591/jpit.v11i1.10021.

[23] Ari Muzakir and Uci Suriani, "Model Deteksi Berita Palsu Menggunakan Pendekatan Bidirectional Long Short Term Memory (BiLSTM)," *Journal of Computer and Information Systems Ampera*, vol. 4, no. 2, 2023.

[24] H. Bhagwani, S. Agarwal, A. Kodipalli, and R. J. Martis, "Targeting class imbalance problem using GAN," in *2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*, IEEE, Dec. 2021, pp. 318–322. doi: 10.1109/ICEECCOT52851.2021.9708011.

[25] H. Yang and Y. Zhou, "IDA-GAN: A Novel Imbalanced Data Augmentation GAN," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, Jan. 2021, pp. 8299–8305. doi: 10.1109/ICPR48806.2021.9411996.