

Analisis Sentimen Kasus Korupsi Pertamina Dengan Naive Bayes Dan SMOTE

Paundra Nabiilah M.H⁻¹, Ionia Veritawati⁻²

Jurusan Teknik Informatika, Fakultas Teknik, Universitas Pancasila⁻¹

email : paundraharahap@gmail.com⁻¹, ionia.veritawati@univpancasila.ac.id⁻²

(Naskah masuk: 15 April 2026; diterima untuk diterbitkan: 10 Mei 2026)

Abstract— Kasus dugaan korupsi PT Pertamina (Persero) memicu respons luas masyarakat di media sosial yang menghasilkan data opini dalam jumlah besar dan bersifat tidak terstruktur. Penelitian ini bertujuan untuk menganalisis sentimen publik terhadap kasus tersebut menggunakan pendekatan klasifikasi teks berbasis pembelajaran mesin. Data dikumpulkan dari Platform X (Twitter) berdasarkan kata kunci terkait korupsi Pertamina dan diproses melalui tahapan *preprocessing* yang meliputi pembersihan teks, normalisasi, penghapusan *stopword*, dan *stemming*. Pelabelan sentimen dilakukan secara otomatis menggunakan model Indonesian *RoBERTa Base Sentiment Classifier* untuk menghasilkan tiga kategori sentimen, yaitu positif, netral, dan negatif. Representasi fitur menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF), sedangkan proses klasifikasi dilakukan dengan algoritma *Naive Bayes*. Untuk mengatasi ketidakseimbangan distribusi kelas, diterapkan metode *Synthetic Minority Over-sampling Technique* (SMOTE) pada data latih. Evaluasi model dilakukan dengan *5-Fold Cross Validation* dan *split validation* dengan metrik akurasi, *precision*, *recall*, *F1-score*, serta analisis ROC-AUC. Hasil penelitian menunjukkan bahwa model awal tanpa penyeimbangan menghasilkan rata-rata akurasi sebesar 45,1%, sedangkan penerapan SMOTE meningkatkan akurasi menjadi 61% pada pengujian *split validation*. Peningkatan performa terutama terlihat pada kelas sentimen positif yang sebelumnya memiliki tingkat *recall* rendah. Temuan ini menunjukkan bahwa model dengan penanganan ketidakseimbangan data berperan penting dalam meningkatkan kinerja klasifikasi sentimen pada data media sosial berbahasa Indonesia yang bertema isu sosial-politik.

Kata Kunci : analisis sentimen, korupsi Pertamina, Naive Bayes, SMOTE, TF-IDF, Indonesian RoBERTa

1. PENDAHULUAN

Perkembangan media sosial telah mengubah cara masyarakat menyampaikan opini terhadap isu-isu publik, termasuk kasus dugaan korupsi yang melibatkan institusi strategis negara. Platform seperti X (Twitter) memungkinkan penyebaran opini secara cepat dan masif, sehingga membentuk ruang diskursus digital yang dinamis [1]. Kasus dugaan korupsi PT Pertamina (Persero) menjadi salah satu isu yang memicu respons luas dari masyarakat, menghasilkan ribuan unggahan yang mencerminkan persepsi, kritik, maupun dukungan

terhadap berbagai pihak yang terlibat [2]. Volume data yang besar dan bersifat tidak terstruktur tersebut memerlukan pendekatan komputasional agar dapat dianalisis secara sistematis.

Analisis sentimen merupakan salah satu pendekatan dalam *text mining* yang bertujuan untuk mengidentifikasi dan mengklasifikasikan opini menjadi kategori tertentu, seperti positif, netral, dan negatif [3]. Dalam konteks isu sosial-politik, analisis sentimen tidak hanya berfungsi untuk mengetahui kecenderungan opini publik, tetapi juga dapat menjadi indikator persepsi masyarakat terhadap kebijakan atau institusi tertentu. Berbagai penelitian telah memanfaatkan algoritma klasifikasi seperti *Naive Bayes*, *Support Vector Machine*, dan metode berbasis *deep learning* untuk menganalisis sentimen pada data media sosial berbahasa Indonesia. *Naive Bayes* tetap menjadi salah satu metode yang banyak digunakan karena sederhana, efisien, dan memiliki performa yang cukup baik pada data teks berdimensi tinggi [4].

Meskipun demikian, salah satu tantangan utama dalam analisis sentimen pada data media sosial adalah ketidakseimbangan distribusi kelas. Pada isu-isu tertentu, distribusi opini sering kali didominasi oleh satu kelas sentimen, sehingga model klasifikasi cenderung bias terhadap kelas mayoritas. Kondisi ini dapat menyebabkan rendahnya kemampuan model dalam mendeteksi kelas minoritas, terutama pada kategori sentimen positif atau netral yang jumlahnya lebih sedikit [5]. Permasalahan ketidakseimbangan data menjadi krusial karena berdampak langsung pada kualitas evaluasi model dan interpretasi hasil analisis.

Untuk mengatasi permasalahan tersebut, teknik penyeimbangan data seperti *Synthetic Minority Over-sampling Technique* (SMOTE) dapat diterapkan pada data latih. SMOTE bekerja dengan menghasilkan sampel sintesis pada kelas minoritas sehingga distribusi kelas menjadi lebih seimbang [6]. Penerapan teknik ini diharapkan dapat meningkatkan kemampuan model dalam mengenali pola pada kelas yang sebelumnya kurang terwakili. Selain itu, proses pelabelan sentimen secara otomatis menggunakan model berbasis transformer, seperti Indonesian *RoBERTa*, memungkinkan pengolahan data dalam skala besar dengan konsistensi yang

lebih baik dibandingkan pelabelan manual [7].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menganalisis sentimen tanggapan masyarakat terhadap kasus dugaan korupsi PT Pertamina menggunakan algoritma *Naïve Bayes* dengan representasi fitur TF-IDF [8]. Penelitian ini juga mengevaluasi pengaruh penerapan *SMOTE* dalam menangani ketidakseimbangan data terhadap kinerja model klasifikasi. Evaluasi dilakukan menggunakan *5-Fold Cross Validation* dan *split validation* dengan metrik akurasi, *precision*, *recall*, *F1-score*, serta ROC-AUC. Hasil penelitian diharapkan dapat memberikan gambaran yang lebih komprehensif mengenai persepsi publik di media sosial serta menunjukkan pentingnya penanganan ketidakseimbangan data dalam analisis sentimen berbasis pembelajaran mesin.

2. METODE PENELITIAN

Penelitian ini menggunakan data teks tanggapan masyarakat mengenai kasus dugaan korupsi PT Pertamina yang diperoleh dari Platform X (Twitter) sebagai bahan utama analisis. Pendekatan yang digunakan adalah *text mining* dengan kerangka kerja *Knowledge Discovery in Databases* (KDD) untuk memastikan proses pengolahan data dilakukan secara sistematis, mulai dari pemilihan data hingga evaluasi model klasifikasi [9]. Metode klasifikasi yang diterapkan adalah *Naïve Bayes* dengan representasi fitur TF-IDF, serta penanganan ketidakseimbangan data menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE).

2.1. Knowledge Discovery in Databases (KDD)

KDD merupakan proses sistematis untuk mengekstraksi pengetahuan dari data dalam beberapa tahapan terstruktur, yaitu *selection*, *preprocessing*, *transformation*, *data mining*, dan *evaluation*. Kerangka kerja ini digunakan agar proses analisis sentimen dilakukan secara terarah dan terukur. Pada tahap *selection*, data dikumpulkan dari Platform X menggunakan kata kunci yang berkaitan dengan kasus korupsi PT Pertamina. Data yang diperoleh berupa teks unggahan berbahasa Indonesia yang kemudian disimpan dalam format dataset untuk diproses lebih lanjut.

Tahap *preprocessing* dilakukan untuk membersihkan dan menyiapkan data teks sebelum dianalisis. Proses ini meliputi penghapusan URL, *mention*, hashtag, angka, tanda baca, serta karakter non-alfabet. Selain itu dilakukan normalisasi teks, pengubahan huruf menjadi *lowercase*, penghapusan *stopword*, dan proses *stemming* untuk mengubah kata menjadi bentuk dasar. Tahapan ini bertujuan untuk mengurangi *noise* dan meningkatkan kualitas representasi teks. Pada tahap *transformation*, teks yang telah dibersihkan diubah menjadi representasi numerik menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Representasi ini mengubah setiap dokumen menjadi vektor fitur berdasarkan bobot kemunculan kata, sehingga dapat digunakan sebagai masukan dalam proses klasifikasi.

Tahap *data mining* dilakukan dengan menerapkan algoritma *Naïve Bayes* untuk mengklasifikasikan teks ke dalam tiga kategori sentimen, yaitu positif, netral, dan negatif. Sebelum proses klasifikasi, pelabelan sentimen dilakukan

secara otomatis menggunakan model *Indonesian RoBERTa Base Sentiment Classifier* untuk menghasilkan *ground truth* pada dataset. Untuk mengatasi ketidakseimbangan distribusi kelas, metode *SMOTE* diterapkan pada data latih guna menghasilkan sampel sintesis pada kelas minoritas sehingga distribusi data menjadi lebih seimbang [6].

Tahap *evaluation* dilakukan untuk menilai kinerja model klasifikasi. Evaluasi dilakukan menggunakan *5-Fold Cross Validation* dan *split validation*. Metrik yang digunakan meliputi akurasi, *precision*, *recall*, *F1-score*, serta analisis *Receiver Operating Characteristic-Area Under Curve* (ROC-AUC) untuk mengukur kemampuan model dalam membedakan antar kelas sentimen.

2.2. Representasi Fitur TF-IDF

TF-IDF merupakan metode pembobotan kata yang umum digunakan dalam analisis teks [10]. Metode ini mengukur pentingnya suatu kata dalam sebuah dokumen relatif terhadap keseluruhan dokumen dalam korpus. Nilai TF-IDF dihitung berdasarkan frekuensi kemunculan kata dalam dokumen dan frekuensi kemunculan kata tersebut pada seluruh dokumen.

Secara umum, pembobotan TF-IDF dihitung menggunakan persamaan (1).

$$TF-IDF(t, d) = TF(t, d) \times IDF(t) \quad (1)$$

di mana $TF(t, d)$ merupakan frekuensi kemunculan *term* t pada dokumen d , dan $IDF(t)$ dihitung menggunakan persamaan (2).

$$IDF(t) = \log \frac{N}{df(t)} \quad (2)$$

dengan N adalah jumlah seluruh dokumen dan $df(t)$ adalah jumlah dokumen yang mengandung *term* t . Representasi ini menghasilkan vektor numerik berdimensi tinggi yang menjadi input bagi algoritma klasifikasi.

2.3. Algoritma Naïve Bayes

Naïve Bayes merupakan algoritma klasifikasi berbasis probabilistik yang didasarkan pada Teorema Bayes dengan asumsi independensi antar fitur. Dalam konteks analisis teks, algoritma ini menghitung probabilitas suatu dokumen termasuk dalam kelas tertentu berdasarkan distribusi kata pada dokumen tersebut.

Secara umum, probabilitas suatu dokumen d termasuk dalam kelas c dihitung menggunakan persamaan (3).

$$P(c|d) = \frac{P(d|c) P(c)}{P(d)} \quad (3)$$

Karena $P(d)$ bersifat konstan untuk setiap kelas, proses klasifikasi dilakukan dengan memilih kelas dengan nilai probabilitas posterior $P(c|d)$ tertinggi. *Naïve Bayes* dipilih karena efisien dalam menangani data berdimensi tinggi dan memiliki performa yang stabil pada klasifikasi teks.

2.4. Penanganan Ketidakseimbangan Data dengan SMOTE

Distribusi sentimen pada data media sosial sering kali tidak seimbang, di mana salah satu kelas mendominasi jumlah data. Kondisi ini dapat menyebabkan model cenderung bias terhadap kelas mayoritas. Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan *Synthetic Minority Over-sampling Technique* (SMOTE) pada data latih.

SMOTE bekerja dengan menghasilkan sampel sintetis pada kelas minoritas berdasarkan kedekatan antar sampel dalam ruang fitur. Teknik ini membantu meningkatkan representasi kelas minoritas tanpa sekadar menduplikasi data yang sudah ada. Dengan distribusi kelas yang lebih seimbang, model diharapkan mampu mengenali pola pada setiap kategori sentimen secara lebih proporsional.

3. HASIL DAN PEMBAHASAN

3.1. Evaluasi Model Menggunakan 5-Fold Cross Validation

Hasil evaluasi awal menggunakan *5-Fold Cross Validation* ditunjukkan pada Tabel 1. Nilai akurasi pada masing-masing fold adalah 0.419, 0.4413, 0.5447, 0.5307, dan 0.3193 dengan rata-rata akurasi sebesar 0.451. Variasi nilai akurasi antar *fold* menunjukkan bahwa performa model belum stabil dan masih sensitif terhadap pembagian data. *Fold* kelima menghasilkan akurasi paling rendah, yang mengindikasikan adanya ketidakseimbangan distribusi kelas atau perbedaan karakteristik data pada subset tersebut.

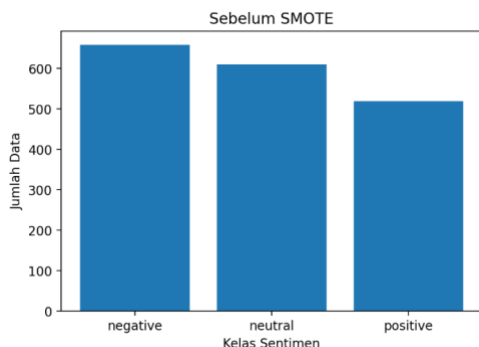
Tabel 1. Hasil Evaluasi *Cross Validation*

Fold	Akurasi
1	0,419
2	0,441
3	0,545
4	0,531
5	0,319
Rata-rata akurasi: 0.451	

Rata-rata akurasi sebesar 45,1% menunjukkan bahwa model tanpa penanganan ketidakseimbangan data belum mampu melakukan klasifikasi secara optimal. Nilai ini masih berada jauh dari batas ideal untuk klasifikasi multi-kelas yang seimbang. Hasil ini memperkuat indikasi bahwa distribusi kelas yang tidak merata berpengaruh terhadap performa model.

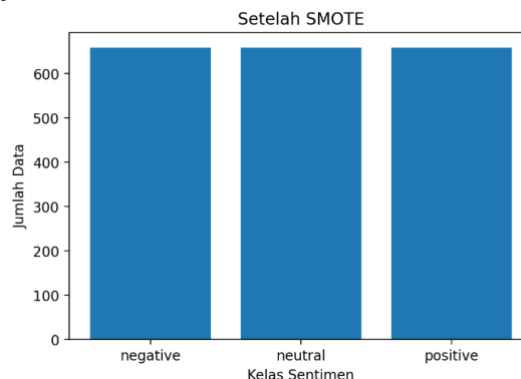
3.2 Distribusi Kelas Sebelum dan Setelah *SMOTE*

Distribusi kelas sebelum dilakukan penyeimbangan data ditunjukkan pada **Gambar 2**. Terlihat bahwa kelas negatif memiliki jumlah data paling besar, diikuti oleh kelas netral, sedangkan kelas positif memiliki jumlah paling sedikit. Ketimpangan ini berpotensi menyebabkan model lebih cenderung memprediksi kelas mayoritas dan mengabaikan pola pada kelas minoritas.



Gambar 2. Distribusi Data Sebelum *SMOTE*

Setelah penerapan *SMOTE*, distribusi kelas menjadi lebih seimbang sebagaimana ditunjukkan pada **Gambar 3**. Jumlah data pada masing-masing kelas relatif setara, sehingga model memiliki kesempatan yang sama untuk mempelajari pola dari setiap kategori sentimen. Penyeimbangan ini diterapkan hanya pada data latih untuk menjaga validitas evaluasi pada data uji.



Gambar 3. Distribusi Data Setelah *SMOTE*

3.3. Hasil Klasifikasi dan Analisis Performa

Hasil klasifikasi pada skenario dengan *SMOTE* ditunjukkan pada **Tabel 2** melalui *Classification Report*. Jumlah data awal 1790 dan dilakukan split 80:20, maka jumlah data yang digunakan untuk data latih adalah 1432 dan untuk data test adalah 358. Model menghasilkan akurasi sebesar 0.61 pada data uji. Untuk kelas negatif, diperoleh *precision* 0.71, *recall* 0.80, dan *F1-score* 0.75. Kinerja pada kelas ini tergolong baik karena *recall* yang tinggi menunjukkan kemampuan model dalam mendeteksi opini negatif secara konsisten.

Tabel 2. *Classification Report*

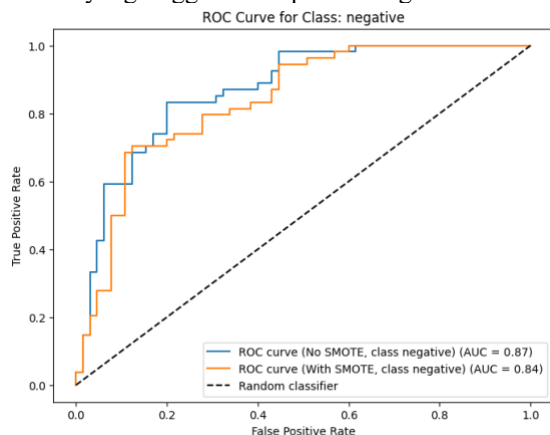
Label	Precision	Recall	F1-score	Support
negative	0.71	0.80	0.75	133
neutral	0.56	0.62	0.59	127
positive	0.46	0.32	0.38	98
accuracy	—	—	0.61	358
macro avg	0.58	0.58	0.57	358
weighted avg	0.59	0.61	0.59	358

Pada kelas netral, diperoleh *precision* 0.56, *recall* 0.62, dan *F1-score* 0.59. Performa ini berada pada tingkat moderat, yang menunjukkan bahwa model masih mengalami kesulitan dalam membedakan opini netral dengan opini negatif atau positif yang memiliki struktur kalimat serupa.

Kelas positif menunjukkan performa paling rendah dengan *precision* 0.46, *recall* 0.32, dan *F1-score* 0.38. Rendahnya *recall* mengindikasikan bahwa sebagian besar opini positif masih salah diklasifikasikan sebagai kelas lain. Hal ini dapat disebabkan oleh jumlah data positif yang lebih sedikit serta karakteristik teks positif pada isu sosial-politik yang cenderung implisit atau mengandung ironi.

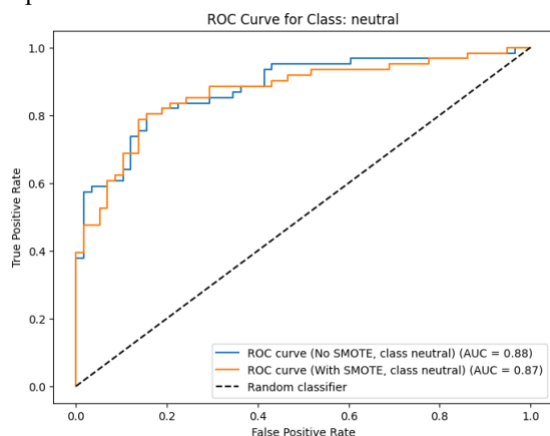
3.4 Analisis ROC dan Perbandingan Skenario

Perbandingan kurva ROC untuk kelas negatif ditunjukkan pada **Gambar 4**. Nilai AUC tanpa *SMOTE* sebesar 0.87, sedangkan dengan *SMOTE* sebesar 0.84. Meskipun terjadi sedikit penurunan AUC, model tetap menunjukkan kemampuan diskriminasi yang tinggi terhadap kelas negatif.



Gambar 4. Kurva ROC Kelas Negatif

Untuk kelas netral, hasil pada **Gambar 5** menunjukkan AUC sebesar 0.88 tanpa *SMOTE* dan 0.87 dengan *SMOTE*. Perubahan yang terjadi relatif kecil dan menunjukkan bahwa penyeimbangan data tidak secara signifikan memengaruhi kemampuan model dalam membedakan kelas netral.

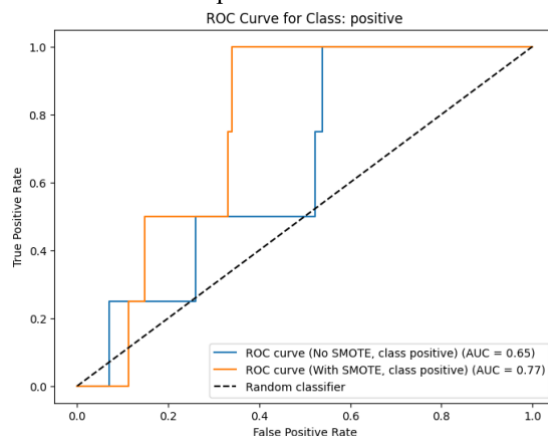


Gambar 5. Kurva ROC Kelas Netral

Perubahan paling signifikan terlihat pada kelas positif sebagaimana ditunjukkan pada **Gambar 6**. Nilai AUC meningkat dari 0.65 tanpa *SMOTE* menjadi 0.77 setelah penerapan *SMOTE*. Peningkatan ini menunjukkan bahwa penyeimbangan data memberikan dampak nyata terhadap kemampuan model dalam mengenali pola pada kelas minoritas. Dengan distribusi data yang lebih seimbang, model menjadi lebih sensitif terhadap karakteristik teks positif yang sebelumnya kurang terwakili.

Secara keseluruhan, hasil eksperimen menunjukkan bahwa ketidakseimbangan distribusi kelas berpengaruh terhadap performa model klasifikasi sentimen. Penerapan *SMOTE* mampu meningkatkan kemampuan model dalam mendeteksi kelas positif, meskipun pada beberapa kelas terjadi

sedikit penurunan nilai AUC. Temuan ini menegaskan bahwa strategi penanganan imbalance menjadi faktor penting dalam analisis sentimen berbasis pembelajaran mesin pada data media sosial bertema isu sosial-politik.



Gambar 6. Kurva ROC Kelas Positif

Hal tersebut juga diperkuat oleh hasil evaluasi *5-Fold Cross Validation* yang disajikan pada **Tabel 3**. Sebelum penerapan *SMOTE*, nilai akurasi pada masing-masing *fold* menunjukkan fluktuasi yang cukup besar, dengan rentang antara 0,319 hingga 0,545 dan rata-rata sebesar 0,451. Variasi ini mengindikasikan bahwa model belum memiliki konsistensi performa yang baik dan cenderung terpengaruh oleh distribusi data yang tidak seimbang.

Tabel 3. Perbandingan Akurasi *5-Fold Cross Validation* Sebelum dan Sesudah *SMOTE*

<i>Fold</i>	Before <i>SMOTE</i>	After <i>SMOTE</i>
<i>Fold 1</i>	0,419	0,58
<i>Fold 2</i>	0,441	0,60
<i>Fold 3</i>	0,545	0,63
<i>Fold 4</i>	0,531	0,62
<i>Fold 5</i>	0,319	0,61
Rata-rata	0,451	0,61

Setelah dilakukan penyeimbangan data menggunakan *SMOTE*, seluruh *fold* mengalami peningkatan akurasi secara konsisten, dengan rentang nilai antara 0,58 hingga 0,63 dan rata-rata meningkat menjadi 0,61. Selain kenaikan rata-rata akurasi sebesar 0,159 poin, variasi antar *fold* juga menjadi lebih kecil dibandingkan sebelum *SMOTE*, yang menunjukkan bahwa model menjadi lebih stabil dan memiliki kemampuan generalisasi yang lebih baik. Dengan demikian, hasil pada Tabel 3 semakin menegaskan bahwa penerapan *SMOTE* tidak hanya meningkatkan performa secara kuantitatif, tetapi juga memperbaiki konsistensi model dalam menghadapi variasi pembagian data.

4. SIMPULAN

Berdasarkan hasil penelitian, model analisis sentimen menggunakan algoritma *Naïve Bayes* mampu mengklasifikasikan tanggapan masyarakat terhadap kasus dugaan korupsi PT Pertamina ke dalam tiga kategori sentimen, yaitu negatif, netral, dan positif. Evaluasi menggunakan *5-Fold Cross Validation* menunjukkan rata-rata akurasi sebesar 45,1% pada data tanpa penyeimbangan kelas, yang mengindikasikan bahwa ketidakseimbangan distribusi sentimen memengaruhi stabilitas dan performa model.

Penerapan *Synthetic Minority Over-sampling Technique* pada data latih meningkatkan performa klasifikasi, dengan akurasi mencapai 61% pada pengujian *split validation*. Peningkatan paling signifikan terlihat pada kelas sentimen positif, yang sebelumnya memiliki kemampuan deteksi rendah. Hasil analisis ROC juga menunjukkan bahwa penanganan ketidakseimbangan data berkontribusi terhadap peningkatan kemampuan model dalam membedakan kelas minoritas.

Temuan ini menunjukkan bahwa strategi penanganan ketidakseimbangan data memiliki peran penting dalam analisis sentimen berbasis pembelajaran mesin, khususnya pada isu sosial-politik di media sosial yang cenderung memiliki distribusi opini tidak merata. Pendekatan kombinasi pelabelan otomatis menggunakan *Indonesian RoBERTa*, representasi fitur TF-IDF, serta klasifikasi *Naïve Bayes* dengan *SMOTE* memberikan kerangka kerja yang sistematis untuk menganalisis persepsi publik secara komputasional.

Penelitian selanjutnya dapat mempertimbangkan penggunaan algoritma klasifikasi lain seperti *Support Vector Machine*, *Random Forest*, atau pendekatan berbasis deep learning untuk membandingkan performa model pada dataset yang sama. Selain itu, analisis dapat diperluas dengan melakukan validasi manual terhadap sebagian data untuk mengukur kualitas pelabelan otomatis yang dihasilkan oleh model transformer.

Pengembangan penelitian juga dapat diarahkan pada analisis temporal untuk melihat dinamika perubahan sentimen publik dari waktu ke waktu, serta eksplorasi teknik penanganan imbalance lain seperti *cost-sensitive learning*. Dengan pengembangan tersebut, analisis sentimen terhadap isu publik di media sosial dapat menghasilkan pemahaman yang lebih komprehensif dan akurat.

PENGAKUAN

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada semua pihak yang telah memberikan dukungan dalam penyelesaian penelitian ini. Penulis juga menyampaikan apresiasi kepada Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pancasila, atas dukungan akademik dan fasilitas yang diberikan selama proses penelitian berlangsung. Selain itu, penulis berterima kasih kepada keluarga dan rekan-rekan yang telah memberikan dukungan moral dan motivasi sehingga penelitian ini dapat terselesaikan dengan lancar. Semoga hasil penelitian ini dapat memberikan kontribusi ilmiah, khususnya dalam pengembangan analisis sentimen berbasis pembelajaran mesin pada data media sosial berbahasa Indonesia.

DAFTAR PUSTAKA

- [1] Reuters, "Indonesia's Pertamina pledges to improve transparency following corruption allegations." Mar. 2025.
- [2] Tempo.co, "Perjalanan Kasus Dugaan Korupsi Pertamina, Kejagung Periksa Sejumlah Saksi." 2025.
- [3] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.
- [4] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," in *AAAI Workshop on Learning for Text Categorization*, 1998.
- [5] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.
- [6] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
- [7] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach." 2019.
- [8] G. Karak, S. Mishra, A. Bandyopadhyay, P. R. S. Rohith, and H. Rathore, "Sentiment Analysis of IMDb Movie Reviews: A Comparative Analysis of Feature Selection and Feature Extraction Techniques BT - Hybrid Intelligent Systems," 2022, pp. 283–294.
- [9] A. Azevedo and M. F. Santos, "KDD , SEMMA AND CRISP-DM : A PARALLEL OVERVIEW Ana Azevedo and M . F . Santos," *IADIS Eur. Conf. Data Min.*, no. January 2008, pp. 182–185, 2008, [Online]. Available: <http://recipp.ipp.pt/handle/10400.22/136%0Ahttp://recipp.ipp.pt/bitstream/10400.22/136/3/KDD-CRISP-SEMMA.pdf>
- [10] K. Agarwal, "A Review on Feature Selection Techniques for Sentiment Analysis," *Interantional J. Sci. Res. Eng. Manag.*, vol. 06, no. 03, pp. 1–15, 2022, doi: 10.55041/ijrsrem12024.