Jurnal Pendidikan dan Teknologi Indonesia (JPTI)
Vol. 5, No. 4, April 2025, Hal. 999-1010
DOI: https://doi.org/10.52436/1.jpti.741
p-ISSN: 2775-4227
e-ISSN: 2775-4219

Optimasi Analisis Sentimen Lowongan Kerja di Twitter Dengan XGBoost-Vader dan Evaluasi SMOTE Borderline

Luthfi Nur Ja'far[*1], L. Budi Handoko[2]

[1,2]Teknik Informatika, Universitas Dian Nuswantoro, Indonesia
Email: [1]111202113633@mhs.dinus.ac.id, [2]handoko@dsn.dinus.ac.id

Abstrak

Perkembangan komunikasi digital telah menjadikan Twitter sebagai platform utama dalam rekrutmen di Indonesia. Namun, analisis sentimen pada platform ini masih jarang diterapkan secara optimal, padahal dapat memberikan wawasan penting bagi pencari kerja dan perekrut dalam memahami persepsi publik terhadap lowongan kerja. Penelitian ini mengembangkan model analisis sentimen menggunakan XGBoost dan VADER untuk mengklasifikasikan postingan lowongan kerja berbahasa Indonesia ke dalam tiga kategori: positif, negatif, dan netral. Dataset terdiri dari 2.181 postingan, dengan rincian 1.711 netral, 414 positif, dan 56 negatif. Untuk menangani ketidakseimbangan data, diterapkan Synthetic Minority Over-sampling Technique (SMOTE) Borderline, yaitu teknik penyeimbangan data yang secara selektif menghasilkan sampel sintetis pada batas keputusan. Namun, teknik ini tidak meningkatkan akurasi model secara signifikan. Sebelum tuning, akurasi model konsisten di 99,95% hingga 100%, sementara setelah tuning, akurasi awalnya sedikit lebih rendah tetapi kemudian stabil di 100%. Evaluasi menggunakan classification report, confusion matrix, dan Stratified K-Fold Cross Validation menunjukkan bahwa model tetap stabil dan mampu menggeneralisasi data dengan baik tanpa indikasi overfitting. Dibandingkan pendekatan sebelumnya, penelitian ini menunjukkan bahwa kombinasi XGBoost dan VADER tanpa balancing data tambahan tetap mampu memberikan analisis sentimen yang lebih akurat dan stabil untuk platform lowongan kerja di Indonesia. Hasil ini berkontribusi dalam pengembangan model analisis sentimen berbasis machine learning yang lebih sesuai dengan karakteristik bahasa Indonesia, serta membuka peluang penelitian lebih lanjut dalam analisis opini publik di media sosial.

Kata kunci: *Analisis Sentimen, Lowongan Kerja, SMOTE Borderline, VADER, XGBoost.*

Optimizing Twitter Job Sentiment Analysis with XGBoost-Vader and SMOTE Borderline Evaluation

Abstract

The development of digital communication has made Twitter a primary platform for recruitment in Indonesia. However, sentiment analysis on this platform is still rarely applied optimally, even though it can provide valuable insights for job seekers and recruiters in understanding public perceptions of job vacancies. This study develops a sentiment analysis model using XGBoost and VADER to classify Indonesian-language job postings into three categories: positive, negative, and neutral. The dataset consists of 2,181 posts, with 1,711 classified as neutral, 414 as positive, and 56 as negative. To address data imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) Borderline was applied, a data balancing technique that selectively generates synthetic samples at the decision boundary. However, this technique did not significantly improve the model's accuracy. Before tuning, the model's accuracy consistently ranged from 99.95% to 100%, while after tuning, the initial accuracy was slightly lower but later stabilized at 100%. Evaluation using a classification report, confusion matrix, and Stratified K-Fold Cross Validation demonstrated that the model remained stable and effectively generalized the data without indications of overfitting. Compared to previous approaches, this study shows that the combination of XGBoost and VADER without additional data balancing can still provide more accurate and stable sentiment analysis for job vacancy platforms in Indonesia. These findings contribute to the development of machine learning-based sentiment analysis models better suited to the characteristics of the Indonesian language and open opportunities for further research in public opinion analysis on social media.

Keywords: *Sentiment Analysis, Job Vacancy, SMOTE Borderline, VADER, XGBoost.*

p-ISSN: 2775-4227
e-ISSN: 2775-4219

1. PENDAHULUAN

Transformasi digital telah memberikan dampak signifikan terhadap mekanisme distribusi informasi. Kemajuan teknologi informasi tidak hanya mengubah cara informasi disebarluaskan tetapi juga memengaruhi strategi pemasaran secara global, terutama melalui pemanfaatan platform media sosial yang semakin luas. Selain berfungsi sebagai media komunikasi dan sumber informasi, media sosial kini memainkan peran penting dalam sektor bisnis, mencakup strategi promosi, kolaborasi dalam pengembangan produk, serta menjadi salah satu sumber utama informasi bagi masyarakat, termasuk dalam pencarian lowongan kerja [1]. Dalam era digital, analisis sentimen semakin banyak digunakan untuk memahami persepsi publik terhadap lowongan kerja dan strategi rekrutmen. Analisis sentimen merupakan proses untuk mengidentifikasi dan menginterpretasikan opini masyarakat terhadap suatu topik tertentu. Dalam konteks ini, penelitian dilakukan untuk mengetahui pendapat masyarakat terhadap informasi lowongan pekerjaan yang tersebar di media sosial Twitter [2]. Di antara berbagai platform media sosial, Twitter merupakan salah satu yang paling populer di Indonesia, dengan tren penggunaan yang terus meningkat. Hal ini mencerminkan semakin luasnya adopsi teknologi dalam proses rekrutmen digital [3]. Selain menjadi sarana komunikasi, Twitter juga menjadi sumber data bagi berbagai penelitian, termasuk analisis sentimen dalam *opinion mining* yang digunakan untuk memahami persepsi publik terhadap suatu isu atau tren pasar [4].

Berbagai metode telah digunakan dalam penelitian sebelumnya untuk menganalisis sentimen, baik dengan pendekatan *machine learning* maupun *lexicon-based*. Pratama & Setyaningsih (2023) menggunakan Naïve Bayes untuk menganalisis sentimen lowongan kerja, dengan mayoritas sentimen netral (64%). Vader digunakan untuk pelabelan otomatis dengan akurasi 98%, tetapi terbatas pada satu metode. Hidayat & Sugiyono (2023) membandingkan Naïve Bayes dan SVM dalam analisis sentimen perekrutan PPPK, menghasilkan akurasi masing-masing 96,14% dan 94,80%, namun belum mengeksplorasi pendekatan hybrid. Asyaroh & Fitriani (2023) menerapkan Random Forest dan XGBoost untuk menganalisis sentimen kekerasan dalam rumah tangga, dengan akurasi XGBoost mencapai 86%, tetapi mengalami overfitting dengan selisih 12% antara pelatihan dan pengujian. Widiarta et al. (2023) menggunakan XGBoost dalam analisis sentimen kebijakan PPKM, sementara Yulistiani dan Styawati (2024) menerapkannya dalam analisis sentimen calon presiden 2024 Hasilnya menunjukkan akurasi masing-masing 85,27% dan 96% setelah augmentasi dataset dan *stemming*, tetapi metode leksikon belum digunakan. Nurkarifin et al. (2024) meneliti sentimen pengguna aplikasi lowongan kerja dengan *Lexicon-Based Features* dan SVM, menghasilkan akurasi 76% hingga 82%, menyoroti pentingnya penyempurnaan kamus leksikon.

Penelitian-penelitian terdahulu cenderung berfokus pada metode individual seperti Naïve Bayes, SVM, atau Random Forest tanpa mempertimbangkan pendekatan *hybrid* untuk meningkatkan performa model. Selain itu, sebagian besar penelitian menggunakan dataset yang tidak seimbang tanpa eksplorasi mendalam terhadap Teknik *oversampling* untuk penanganan ketidakseimbagan data [5]. Meskipun beberapa studi telah menerapkan augmentasi data, hasilnya menunjukkan keterbatasan dalam mitigasi *overfitting* dan peningkatan akurasi [2], [4], [6]. Belum ada penelitian yang secara spesifik mengalisis sentimen lowongan kerja di Twitter menggunakan pendekatan hybrid XGBoost-Vader. Selain itu, efektivitas SMOTE Borderline dalam menangani ketidakseimbangan data pada analisis sentimen lowongan kerja masih jarang dieksplorasi [7]. Hal ini mendorong peneliti untuk melakukan penelitian dengan judul "*Optimasi Analisis Sentimen Lowongan Kerja di Twitter Dengan XGBoost-Vader dan Evaluasi SMOTE Borderline*".

Penelitian ini bertujuan untuk mengevaluasi efektivitas pendekatan hybrid XGBoost-Vader dalam analisis sentiment lowongan kerja di Twitter. Selain itu, penelitian ini membandingkan performa model pada data seimbang dan tidak seimbang menggunakan teknik *oversampling* SMOTE Borderline serta menganalisis dampak *tuning hyperparameter* terhadap akurasi dan potensi *overfitting* [8], [9], [10]. Hasil penelitian ini diharapkan dapat menjadi acuan dalam meningkatkan akurasi model analisis sentimen lowongan kerja serta memberikan wawasan lebih lanjut mengenai efektivitas metode *balancing* data dalam konteks yang serupa [11].

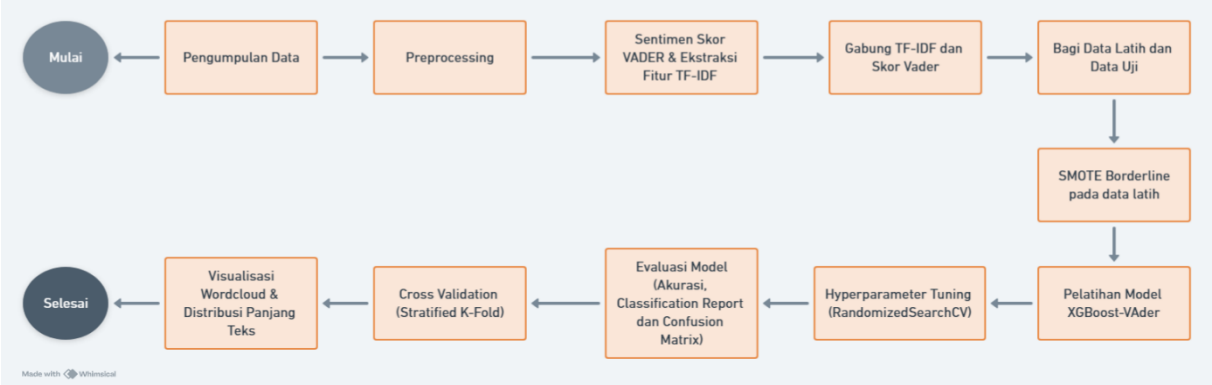
2. METODE PENELITIAN

2.1 Alur Penelitian

Gambar 1 menunjukan *flowchart* penelitian yang menggambarkan tahapan penelitian dari pengumpulan data hingga evaluasi model. *Flowchart* penelitian diawali dengan pengumpulan data dari Twitter. Setelah data terkumpul, dilakukan preprocessing yang mencakup pembersihan teks, penghapusan *stopwords*, tokenisasi, *stemming* dan *lemmatization*. Selanjutnya, fitur VADER dan TF-IDF diekstraksi menggunakan NLTK dan *sklearn.features_extraction.text*, kemudian data dibagi menjadi set latih dan set uji. Untuk mengatasi ketidakseimbangan data, diterapkan SMOTE Borderline menggunakan pustaka *imbalanced-learn*, sehingga jumlah sampel setiap kelas menjadi seimbang. Model XGBoost-Vader kemudian dilatih dan dioptimalkan

p-ISSN: 2775-4227
e-ISSN: 2775-4219

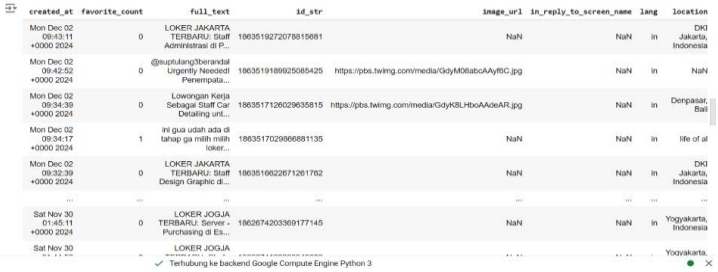
melalui *tuning hyperparameter* menggunakan *GridSearchCV*. Evaluasi dilakukan dengan akurasi, *classification report*, *confusion matrix*, serta *Stratified K-Fold Cross Validation* untuk memastikan model dapat bekerja dengan baik pada distribusi data yang berbeda [17]. Selain itu, *wordcloud* dan distribusi panjang teks divisualisasikan untuk memahami karakteristik data lebih dalam.



Gambar 1 *Flowchart* Alur Proses Penelitian

2.2 Pengumpulan Data

Dataset dalam penelitian ini diperoleh dari media social Twitter dengan menggunakan Tweet Harvest sebagai alat *scraping*. Pemilihan platform ini didasarkan pada tingginya penggunaan Twitter oleh pencari kerja dan perusahaan dalam berbagi informasi terkait lowongan pekerjaan, serta aksesibilitasnya melalui API yang memungkinkan pengambilan data secara sistematis [12]. Proses pengambilan data diawali dengan autentikasi API guna memastikan akses yang sah terhadap data publik. Selanjutnya, dilakukan penyaringan data berdasarkan kata kunci seperti *"lowongan kerja"*, *"loker"*, dan *"karir"*, sehingga hanya unggahan yang relevan dengan penelitian ini yang dikumpulkan. Data dikumpulkan dalam rentang waktu Januari–Desember 2024, dengan tujuan untuk memperoleh representasi tren tahunan dalam pencarian kerja di Indonesia, sehingga mencakup berbagai tren dan opini publik terkait lowongan kerja sepanjang satu tahun penuh.



	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen_name	lang	location
	Mon Dec 02 09:43:11 +0000 2024	0	LOKER JAKARTA TERBARU: Staff Administrasi di P...	1863519272078815681	NaN	NaN	in	DKI Jakarta, Indonesia
	Mon Dec 02 09:42:52 +0000 2024	0	@suptulang3berandal Urgently Needed! Penempata...	1863519189925085425	https://pbs.twimg.com/media/GdyM06abcAAyf6C.jpg	NaN	in	NaN
	Mon Dec 02 09:34:39 +0000 2024	0	Lowongan Kerja Sebagai Staff Car Detailing unt...	1863517126029635815	https://pbs.twimg.com/media/GdyK8LHboAAdeAR.jpg	NaN	in	Denpasar, Bali
	Mon Dec 02 09:34:17 +0000 2024	1	ini gua udah ada di tahap ga milih milih loker...	1863517029866881135	NaN	NaN	in	life of al
	Mon Dec 02 09:32:39 +0000 2024	0	LOKER JAKARTA TERBARU: Staff Design Graphic di...	1863516622671261762	NaN	NaN	in	DKI Jakarta, Indonesia
	...	...	...	...	...	...	...	...
	Sat Nov 30 01:45:11 +0000 2024	0	LOKER JOGJA TERBARU: Server - Purchasing di Es...	1862674203369177145	NaN	NaN	in	Yogyakarta, Indonesia
	Sat Nov 30		LOKER JOGJA					Yogyakarta,

Terhubung ke backend Google Compute Engine Python 3

Gambar 2 *Scraping Data*

Gambar 2 menunjukkan hasil output proses *scraping* data dari Twitter. Dataset yang digunakan terdiri dari 2.181 unggahan dengan distribusi sentimen yang tidak seimbang, yaitu 1.711 tweet netral (78,45%), 414 tweet positif (18,98%), dan 56 tweet negatif (2,57%). Ketidakseimbangan ini dapat menyebabkan model lebih cenderung mengklasifikasikan data ke kelas mayoritas (netral). Untuk mengatasi hal ini, diterapkan SMOTE Borderline guna menyeimbangkan jumlah data di setiap kelas melalui proses *oversampling* [4], [9]. Dengan pendekatan ini, model diharapkan dapat mengenali sentimen dengan lebih akurat serta memberikan wawasan mengenai sentimen publik terhadap informasi lowongan kerja.

2.3 Praproses Data

Tahapan ini bertujuan untuk mempersiapkan data agar siap dianalisis oleh model. Langkah-langkahnya meliputi:

1. Pembersihan Teks (*Cleaning*): Menghapus URL, simbol, angka, dan tanda baca yang tidak relevan dan mengubah teks menjadi huruf kecil untuk konsistensi.
2. *Stopwords Removal*: Menghapus kata-kata umum yang tidak memiliki makna signifikan.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

3. Lemmatization: Mengubah kata menjadi bentuk dasarnya.
4. Sentimen Label dan *Label Encode*: Klasifikasi sentimen teks menjadi tiga kategori positif, netral, dan negatif.

Setelah data melewati tahap praproses untuk memastikan kualitas dan konsistensi, langkah berikutnya adalah membagi dataset menggunakan *train_test_split* dengan 80% untuk data latih dan 20% untuk data uji. Pembagian ini dilakukan secara stratifikasi untuk menjaga proporsi kelas dalam setiap subset [13]. Selanjutnya, dilakukan penghitungan skor sentimen menggunakan metode Vader, yang kemudian dikombinasikan dengan fitur TF-IDF sebelum digunakan dalam proses pelatihan model.

2.1 Pelatihan Model

1. *VADER (Valence Aware Dictionary and Sentiment Reasoner)*

VADER adalah metode berbasis leksikon yang menghitung skor sentimen berdasarkan kata-kata dalam teks. Setiap kata dibandingkan dengan kamus sentimen yang memiliki nilai positif, netral, atau negatif. Berbeda dengan Naïve Bayes**,** yang menggunakan pendekatan probabilistik berdasarkan frekuensi kata dalam dataset latih**,** VADER mampu mengenali sentimen secara langsung tanpa pelatihan model [15]**.** Keunggulan VADER terletak pada kemampuannya menganalisis teks pendek seperti tweet**,** serta mempertimbangkan intensitas emosi melalui huruf kapital, tanda baca, emoji, dan kata penguat (*very*, *extremely*). Sebaliknya, Naïve Bayes sering kesulitan menangkap konteks sentimen dalam kalimat yang mengandung sarkasme atau negasi tidak eksplisit.

Rumus dasar untuk menghitung skor sentimen menggunakan VADER adalah:

$$Sentimen\ Score = \sum_{i=1}^{n} \quad valence\ (w_i) \qquad (1)$$

Secara keseluruhan, skor sentiment dihitung dengan menjumlahkan nilai *valence* untuk setiap kata yang ada dalam teks. Skor kemudian akan menunjukkan apakah teks tersebut memiliki sentiment positif, netral, atau negatif. Berikut gambar 3 merupakan hasil output dari perhitungan skor vader:

	cleaned_text	vader_score
0	loker jogja terbaru content creator cross bord...	0.000
1	loker semarang terbaru cook helper server rack...	0.340
2	kopilimana membuka lowongan kerja posisi super...	0.000
3	trans cargo membuka lowongan kerja posisi crew...	0.000
4	gp mobil membuka lowongan kerja posisi host li...	0.000
..	...	...
195	loker semarang terbaru pelaksana proyek sipil ...	0.000
196	loker semarang terbaru kasir admin accounting ...	0.000
197	loker semarang terbaru barista kopi hati mongi...	0.000
198	loker jakarta terbaru staff it spv purchasing ...	0.128
199	loker jakarta terbaru sales consultant travel ...	0.000

Gambar 3 Perhitungan Skor Vader

2. *XGBoost (Extreme Gradient Boosting)*

XGBoost digunakan dalam penelitian ini untuk klasifikasi sentimen, di mana model ini dilatih untuk memprediksi apakah suatu tweet berisi sentimen positif, netral, atau negatif terkait dengan lowongan kerja. Dibandingkan dengan algoritma lain seperti Naïve Bayes yang mengasumsikan independensi fitur atau SVM yang sensitif terhadap data tidak seimbang, XGBoost menawarkan keunggulan dalam menangani dataset besar dengan fitur yang kompleks serta memiliki mekanisme regularisasi bawaan untuk mengurangi risiko *overfitting* [10]. Rumus dasar untuk fungsi kerugian dalam XGBoost adalah:

$$L(\theta) = \sum_{i=1}^{n} \quad l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \quad \Omega(f_k) \qquad (2)$$

XGBoost menggunakan fungsi kerugian untuk mengukur selisih antara nilai aktual y_i dan prediksi $\hat{y}_i$, dengan mempertimbangkan probabilitas hasil. Regularisasi diterapkan untuk mencegah *overfitting*, dan parameter model θ dioptimalkan selama pelatihan guna meminimalkan fungsi kerugian dan mengontrol kompleksitas model. Dengan teknik boosting, XGBoost secara bertahap memperbaiki kesalahan prediksi melalui pemanfaatan gradien fungsi kerugian di setiap iterasi [8]. Optimasi model dilakukan dengan menggunakan *GridSeacrhCV* untuk mencari kombinasi terbaik dari *hyperparameter* guna menemukan konfigurasi model yang optimal.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

3. *SMOTE Borderline (Synthetic Minority Over-sampling Technique Borderline)*

Teknik SMOTE ini digunakan untuk menangani masalah kelas tidak seimbang dalam dataset, dimana kelas minoritas memiliki lebih sedikit data. SMOTE Borderline secara khusus bekerja pada titik data yang sulit dipisahkan. Metode ini lebih efektif dibandingkan SMOTE biasa dalam menangani data yang berada di batas keputusan. Karena fokusnya pada sampel yang paling sulit diklasifikan. Dibandingkan dengan Random Oversampling, SMOTE Borderline menghasilkan data sintetis yang lebih representatif terhadap pola aslinya.

Rumus untuk menghasilkan titik data sintetis menggunakan SMOTE adalah:

$$X_{new} = X_i + \lambda (X_k - X_i) \qquad (3)$$

SMOTE Borderline menghasilkan data sintetis baru (X_{new}) melalui interpolasi antara data kelas minoritas (X_i) dan tetangga terdekatnya (X_k) menggunakan nilai acak (λ). Teknik ini berfokus pada data di batas keputusan antara kelas, membantu model belajar dari data yang sulit dipisahkan dan meningkatkan performa pada dataset tidak seimbang [9].

2.2 Evaluasi Performa Model

Evaluasi dilakukan menggunakan metrik akurasi, *classification report, confusion matrix*, dan *Stratified 5-Fold Cross Validation* untuk memastikan kestabilan model. Selain itu, *precision, recall,* dan *F1-score* dihitung untuk mengukur keseimbangan prediksi, serta dilakukan perbandingan sebelum dan sesudah penerapan SMOTE Borderline guna menilai peningkatan performa model. Evaluasi model dilakukan berdasarkan metrik berikut:

1. *Precision, Recall,* dan *F1-Score*
   a. *Precision*: Mengukur ketepatan prediksi positif, dihitung sebagai:
   $$Precision = \frac{TP}{TP+FP} \qquad (4)$$
   Semakin tinggi precision, semakin sedikit prediksi positif yang salah.
   b. *Recall*: Mengukur kemampuan model menangkap semua data positif, dihitung sebagai:
   $$Recall = \frac{TP}{TP+FN} \qquad (5)$$
   Recall tinggi berarti model tidak banyak melewatkan data positif.
   c. *F1-Score*: Rata-rata harmonic precision dan recall, dihitung sebagai:
   $$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision \times Recall} \qquad (6)$$
   Digunakan untuk keseimbangan anatara keduanya.
2. *Confusion Matrix*: Matriks yang menunjukkan jumlah prediksi benar dan salah untuk setiap kelas, untuk memahami pola kesalahan yang dibuat oleh model.
3. *Cross-Validation (Stratified 5-Fold):* Teknik ini membagi dataset menjadi 5 *fold* dengan distribusi kelas yang seimbang di setiap fold. Model dilatih pada 4 *fold* dan diuji pada 1 *fold***,** kemudian proses ini diulang sebanyak 5 kali. Rata-rata dari metrik evaluasi seperti akurasi, *precision, recall,* dan *F1-score* dihitung untuk memastikan kestabilan performa model [14]. Pendekatan ini membantu mengurangi *overfitting* dan memberikan gambaran yang lebih akurat mengenai kemampuan model dalam mengklasifikasikan data baru.

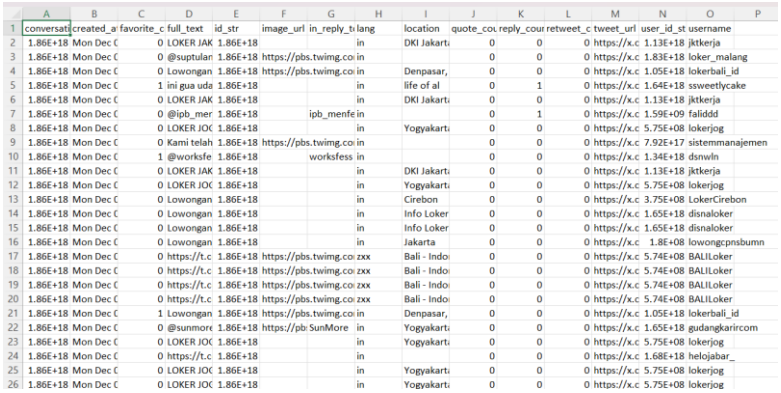
Setelah melalui beberapa tahapan, evaluasi performa model dilakukan untuk menilai efektivitas analisis sentimen. Karena dataset mengalami ketidakseimbangan kelas, evaluasi tidak hanya berfokus pada akurasi, tetapi juga menggunakan precision, recall, dan F1-score untuk mengukur keseimbangan prediksi. Confusion matrix digunakan untuk menganalisis distribusi prediksi, sementara distribusi panjang teks dan *Stratified* 5-Fold Cross-Validation per fold diterapkan untuk memastikan kestabilan model. Model dievaluasi sebelum dan setelah penerapan SMOTE Borderline guna melihat peningkatan dalam menangani ketidakseimbangan data. Perhitungan metrik dilakukan pada dataset uji (20% dari total data) untuk memastikan akurasi model pada data yang belum pernah dilatih.

3. HASIL DAN PEMBAHASAN

3.1 Dataset

Dataset yang digunakan terdiri dari tweet mengenai lowongan kerja yang dikumpulkan menggunakan *tools* Tweet Harvest melalui metode scraping. Data tersebut diambil selama periode Januari hingga Desember 2024 dengan total 2.181 tweet. Dataset hasil pengumpulan disimpan dalam format CSV (*Comma-Separated Values)* menggunakan aplikasi Excel.

p-ISSN: 2775-4227
e-ISSN: 2775-4219



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	conversati	created_a	favorite_c	full_text	id_str	image_url	in_reply_t	lang	location	quote_cou	reply_cou	retweet_c	tweet_url	user_id_st	username	
2	1.86E+18	Mon Dec (	0	LOKER JAK	1.86E+18			in	DKI Jakart	0	0	0	https://x.c	1.13E+18	jktkerja	
3	1.86E+18	Mon Dec (	0	@suptulan	1.86E+18	https://pbs.twimg.co		in		0	0	0	https://x.c	1.83E+18	loker_malang	
4	1.86E+18	Mon Dec (	0	Lowongan	1.86E+18	https://pbs.twimg.co		in	Denpasar,	0	0	0	https://x.c	1.05E+18	lokerbali_id	
5	1.86E+18	Mon Dec (	1	ini gua uda	1.86E+18			in	life of al	0	1	0	https://x.c	1.64E+18	ssweetlycake	
6	1.86E+18	Mon Dec (	0	LOKER JAK	1.86E+18			in	DKI Jakart	0	0	0	https://x.c	1.13E+18	jktkerja	
7	1.86E+18	Mon Dec (	0	@ipb_mer	1.86E+18		ipb_menfe	in		0	1	0	https://x.c	1.59E+09	faliddd	
8	1.86E+18	Mon Dec (	0	LOKER JOC	1.86E+18			in	Yogyakart	0	0	0	https://x.c	5.75E+08	lokerjog	
9	1.86E+18	Mon Dec (	0	Kami telah	1.86E+18	https://pbs.twimg.co		in		0	0	0	https://x.c	7.92E+17	sistemmanajemen	
10	1.86E+18	Mon Dec (	1	@worksfe	1.86E+18		worksfess	in		0	0	0	https://x.c	1.34E+18	dsnwln	
11	1.86E+18	Mon Dec (	0	LOKER JAK	1.86E+18			in	DKI Jakart	0	0	0	https://x.c	1.13E+18	jktkerja	
12	1.86E+18	Mon Dec (	0	LOKER JOC	1.86E+18			in	Yogyakart	0	0	0	https://x.c	5.75E+08	lokerjog	
13	1.86E+18	Mon Dec (	0	Lowongan	1.86E+18			in	Cirebon	0	0	0	https://x.c	3.75E+08	LokerCirebon	
14	1.86E+18	Mon Dec (	0	Lowongan	1.86E+18			in	Info Loker	0	0	0	https://x.c	1.65E+18	disnaloker	
15	1.86E+18	Mon Dec (	0	Lowongan	1.86E+18			in	Info Loker	0	0	0	https://x.c	1.65E+18	disnaloker	
16	1.86E+18	Mon Dec (	0	Lowongan	1.86E+18			in	Jakarta	0	0	0	https://x.c	1.8E+08	lowongcpnsbumn	
17	1.86E+18	Mon Dec (	0	https://t.c	1.86E+18	https://pbs.twimg.co		zxx	Bali - Indo	0	0	0	https://x.c	5.74E+08	BALILoker	
18	1.86E+18	Mon Dec (	0	https://t.c	1.86E+18	https://pbs.twimg.co		zxx	Bali - Indo	0	0	0	https://x.c	5.74E+08	BALILoker	
19	1.86E+18	Mon Dec (	0	https://t.c	1.86E+18	https://pbs.twimg.co		zxx	Bali - Indo	0	0	0	https://x.c	5.74E+08	BALILoker	
20	1.86E+18	Mon Dec (	0	https://t.c	1.86E+18	https://pbs.twimg.co		zxx	Bali - Indo	0	0	0	https://x.c	5.74E+08	BALILoker	
21	1.86E+18	Mon Dec (	1	Lowongan	1.86E+18	https://pbs.twimg.co		in	Denpasar,	0	0	0	https://x.c	1.05E+18	lokerbali_id	
22	1.86E+18	Mon Dec (	0	@sunmore	1.86E+18	https://pb	SunMore	in	Yogyakart	0	0	0	https://x.c	1.65E+18	gudangkarircom	
23	1.86E+18	Mon Dec (	0	LOKER JOC	1.86E+18			in	Yogyakart	0	0	0	https://x.c	5.75E+08	lokerjog	
24	1.86E+18	Mon Dec (	0	https://t.c	1.86E+18			in		0	0	0	https://x.c	1.68E+18	helojabar_	
25	1.86E+18	Mon Dec (	0	LOKER JOC	1.86E+18			in	Yogyakart	0	0	0	https://x.c	5.75E+08	lokerjog	
26	1.86E+18	Mon Dec (	0	LOKER JOC	1.86E+18			in	Yogyakart	0	0	0	https://x.c	5.75E+08	lokerjog	

Gambar 4 Hasil Scraping Data

Gambar 8 menunjukkan hasil pengambilan data (scraping) yang disimpan dalam format CSV. Dataset ini berisi berbagai atribut, termasuk ID percakapan, waktu pembuatan, jumlah suka, teks unggahan, lokasi, jumlah kutipan, balasan, retweet, serta informasi pengguna seperti ID dan username.

3.2 Preprocessing

Selanjutnya, dilakukan tahap preprocessing untuk meningkatkan kualitas data sebelum analisis sentimen dan pelatihan model. Proses ini mencakup pembersihan teks, seperti menghapus URL, karakter non-alfabet, *stopwordss*, serta menerapkan *lemmatization* agar kata berada dalam bentuk dasarnya. Langkah ini penting untuk memastikan data lebih bersih dan siap digunakan dalam proses klasifikasi sentimen serta evaluasi model [13]. Berikut adalah hasil dari masing-masing tahap preprocessing yang telah diterapkan:

1. Pembersihan Teks (*Cleaning*):

```
================================================
        Tabel Teks yang Telah Dibersihkan
================================================


                                          cleaned_text
loker jogja terbaru content creator cross border m
loker semarang terbaru cook helper server racker g
kopilimana membuka lowongan kerja posisi superviso
trans cargo membuka lowongan kerja posisi crew ope
gp mobil membuka lowongan kerja posisi host live p

================================================
```

Gambar 5. Hasil Pembersihan Teks

2. *Stopwords Removal*:
3.

```
================================================
           Tabel Hasil Stopword Removal
================================================


                                       stopword_removed
LOKER JOGJA TERBARU: Content Creator di Cross Bord
LOKER SEMARANG TERBARU: Cook Helper - Server - Rac
Kopilimana membuka lowongan kerja untuk posisi Sup
Trans Cargo membuka lowongan kerja untuk posisi Cr
GP33 Mobil membuka lowongan kerja untuk posisi Hos

================================================
```

Gambar 6. Hasil *Stopwords Removal*

p-ISSN: 2775-4227
e-ISSN: 2775-4219

4. *Lemmatization*:

```
==================================================
        Tabel Teks yang Telah Dilemmatized
==================================================


                                    cleaned_text_lemmatized
loker jogja terbaru content creator cross border m
loker semarang terbaru cook helper server racker g
kopilimana membuka lowongan kerja posisi superviso
trans cargo membuka lowongan kerja posisi crew ope
gp mobil membuka lowongan kerja posisi host live p
```

Gambar 7. *Lemmatization*

5. Sentimen Label dan *Label Encode*:

```
Tabel Sebelum Perubahan:
                                   full_text  sentiment_label
0  LOKER JOGJA TERBARU: Content C                0
1  LOKER SEMARANG TERBARU: Cook H                1
2  Kopilimana saat ini membuka lo                0
3  Trans Cargo saat ini membuka l                0
4  GP33 Mobil saat ini membuka lo                0

Tabel Setelah Perubahan:
                                   full_text  sentiment_label_encoded
0  LOKER JOGJA TERBARU: Content C                        1
1  LOKER SEMARANG TERBARU: Cook H                        2
2  Kopilimana saat ini membuka lo                        1
3  Trans Cargo saat ini membuka l                        1
4  GP33 Mobil saat ini membuka lo                        1
```

Gambar 8. Sentimen Label dan *Label Encode*

3.3 Visualisasi

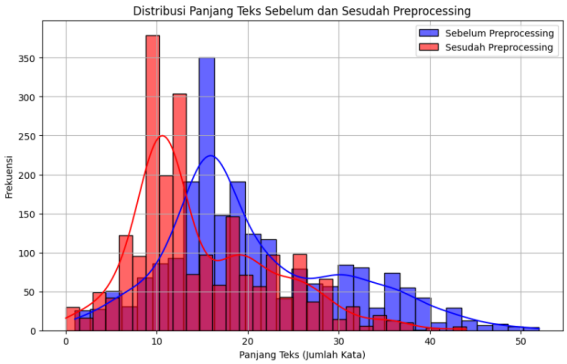
Setelah *preprocessing* selesai, analisis awal dilakukan dengan visualisasi wordcloud untuk memahami distribusi kata dalam dataset. *Wordcloud* ini menampilkan kata-kata dominan berdasarkan sentimen positif, negatif, dan netral, sehingga memberikan gambaran awal tentang pola sentimen dalam data. Proses ini menggunakan *stopwords* untuk menyaring kata-kata umum yang kurang relevan, memastikan hanya kata-kata signifikan yang diutamakan dalam analisis [10]. Berikut adalah hasil *wordcloud* untuk masing-masing sentimen:

Gambar 9. *Wordcloud* Sentimen Negatif, Netral, dan Positif

Gambar 9 ini menampilkan representasi visual dari kata-kata yang paling sering muncul dalam setiap kategori sentimen terkait lowongan kerja di Twitter, yang diklasifikasikan ke dalam tiga kelompok: negatif, netral, dan positif. Pada kategori sentimen negatif, kata-kata yang dominan seperti butuh, freelance, gaji, usaha, dan kerja mengindikasikan adanya keluhan atau ketidakpuasan terhadap kondisi lowongan kerja, seperti kesulitan mendapatkan pekerjaan atau rendahnya kompensasi yang ditawarkan. Sementara itu, kategori sentimen netral didominasi oleh kata-kata seperti lowong kerja, loker baru, buka, info, dan posisi, yang menunjukkan bahwa sebagian besar unggahan dalam kategori ini bersifat informatif dan tidak mengandung emosi tertentu. Di sisi lain, kategori sentimen positif menampilkan kata-kata seperti fresh graduate, teman, follow, dan share, yang mencerminkan optimisme, dukungan, serta ajakan untuk berbagi informasi mengenai peluang kerja. Secara keseluruhan.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Setelah memvisualisasikan distribusi kata-kata, analisis dilanjutkan dengan pemeriksaan distribusi panjang teks untuk memberikan pemahaman yang lebih mendalam mengenai karakteristik tweet yang digunakan dalam penelitian ini [11].

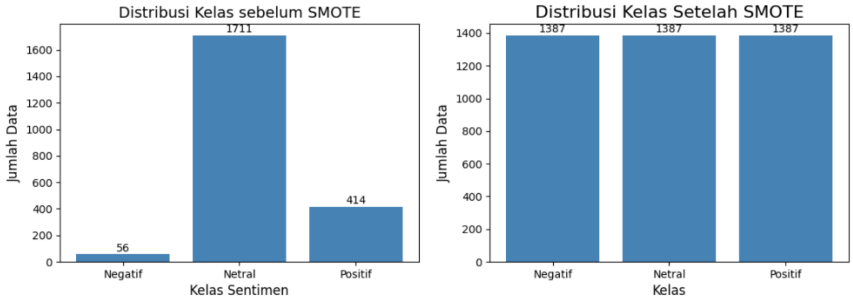


Gambar 10. Distribusi Panjang Teks Sebelum dan Sesudah Preprocessing

Gambar 10 menunjukkan distribusi panjang teks tweet terkait lowongan kerja menunjukkan perubahan setelah preprocessing, dengan rentang awal 1–52 kata dan berkurang menjadi 0–44 kata. Rata-rata panjang teks sebelum preprocessing adalah 20,47 kata (median 18), sedangkan setelah preprocessing turun menjadi 14,88 kata (median 12). Pengurangan ini mencerminkan efektivitas *preprocessing* dalam menyaring kata yang kurang relevan. Analisis ini menunjukkan dampak *preprocessing* terhadap struktur teks dan potensi pengaruhnya pada hasil sentimen. Selain itu, keseimbangan distribusi label sentimen tetap perlu diperhatikan agar model dapat mengenali kelas dengan jumlah sampel lebih sedikit secara optimal.

3.4 SMOTE Borderline

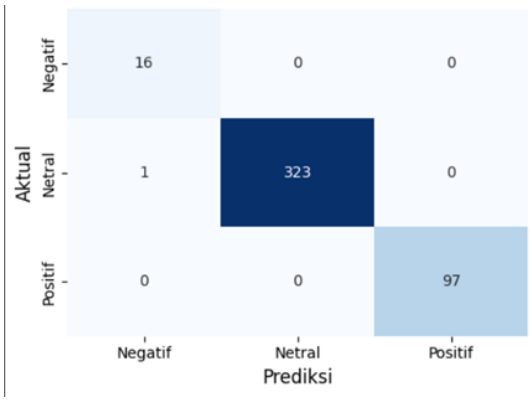
Sebelum diterapkan, dataset menunjukkan dominasi kelas netral yang signifikan, sementara jumlah sampel pada kelas negatif dan positif jauh lebih sedikit. Setelah SMOTE Borderline, distribusi data menjadi lebih seimbang seperti yang ditunjukkan pada gambar berikut:



Gambar 11. Distribusi Label Sentimen Sebelum dan Sesudah *SMOTE Borderline*

Meskipun distribusi data telah diseimbangkan, evaluasi model menunjukkan bahwa SMOTE Borderline tidak memberikan dampak signifikan terhadap performa klasifikasi. Hasil ini menegaskan bahwa model XGBoost-VADER tetap stabil dengan atau tanpa teknik *oversampling*. SMOTE Borderline tidak meningkatkan atau menurunkan akurasi, tetapi hanya berfungsi sebagai langkah *preprocessing* untuk menyeimbangkan jumlah data antar kelas. Untuk memahami lebih lanjut bagaimana model menangani prediksi setelah penerapan SMOTE Borderline, distribusi kesalahan klasifikasi dapat dianalisis melalui *confusion matrix* berikut:

p-ISSN: 2775-4227
e-ISSN: 2775-4219



Gambar 12. *Confusion Matrix*

Gambar 10 menunjukkan bahwa meskipun data telah diseimbangkan dengan SMOTE Borderline, pola prediksi model tetap stabil dengan hanya satu kesalahan klasifikasi pada kelas netral. Ini menegaskan bahwa SMOTE Borderline tidak memengaruhi akurasi, melainkan hanya membantu menyeimbangkan distribusi data.

Tabel 1. Distribusi Prediksi Berdasarkan Confusion Matrix

Kelas	Benar Prediksi (True)	Salah Prediksi (Negatif)	Salah Prediksi (Netral)	Salah Prediksi (Positif)
Sentimen Negatif	16	0	0	0
Sentimen Netral	323	1	0	0
Sentimen Positif	97	0	0	0

Tabel 1 menunjukkan bahwa model XGBoost-VADER hampir sempurna dalam klasifikasi, dengan hanya satu kesalahan pada kelas netral yang diprediksi sebagai negatif. Kelas negatif dan positif diklasifikasikan dengan benar sepenuhnya. Kesalahan ini menunjukkan sedikit sensitivitas model terhadap sentimen negatif, tetapi secara keseluruhan, model tetap stabil dan optimal setelah penerapan SMOTE Borderline dan *tuning hyperparameter*.

3.5 Pelatihan Model XGBoost-VADER

Penelitian ini menggunakan model XGBoost-VADER, yang mengombinasikan XGBoost sebagai model klasifikasi dengan VADER sebagai metode pembobotan sentimen berbasis leksikon. Model ini dilatih menggunakan dataset yang telah diproses, baik tanpa maupun dengan SMOTE Borderline untuk menangani ketidakseimbangan kelas. Pelatihan model dilakukan menggunakan dataset yang telah dibersihkan dan diformat sesuai kebutuhan analisis sentimen. Pembagian data dilakukan dengan rasio 80:20, di mana 80% digunakan untuk pelatihan dan 20% untuk pengujian. Evaluasi model dilakukan berdasarkan metrik akurasi, *precision, recall,* dan *F1-score*, yang dihitung menggunakan *confusion matrix* dari hasil prediksi model.

Untuk meningkatkan performa model, dilakukan *tuning hyperparameter* menggunakan *GridSearchCV* dan diperoleh kombinasi *hyperparameter* terbaik sebagai berikut:

Tabel 2. Hyperparameter Optimal

Hyperparameter	Nilai Optimal	Kegunaan
Subsample	0.8	Mengurangi *overfitting* dengan membatasi jumlah data yang digunakan setiap pohon.
n_estimators	300	Menentukan jumlah pohon dalam model.
max_depth	7	Mengontrol kedalaman pohon untuk mencegah *overfitting*.
learning_rate	0.01	Mengatur kecepatan pembelajaran agar lebih stabil.
colsample_bytree	0.8	Mengurangi *overfitting* dengan membatasi jumlah fitur per pohon.

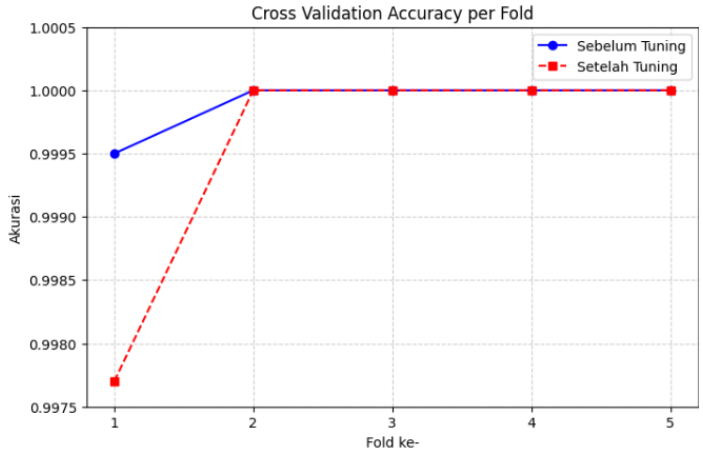
p-ISSN: 2775-4227
e-ISSN: 2775-4219

Tabel 2 menyajikan hasil *tuning hyperparameter* terbaik yang digunakan dalam penelitian ini. Dengan konfigurasi ini, model XGBoost-VADER dioptimalkan untuk mencapai keseimbangan antara akurasi dan generalisasi. Setelah mendapatkan *hyperparameter* terbaik, model diuji kembali pada data uji. Berikut adalah hasil perbandingan sebelum dan sesudah *tuning*:

Tabel 3. Evaluasi Model Sebelum dan Sesudah *Hyperparameter Tuning*

Metrik	Sebelum Tuning (Cross Validation)	Setelah Tuning (Test Set)
Akurasi	0.9995	0.9977
Precision (makro)	0.98	0.98
Recall (makro)	1.00	1.00
F1-score (makro)	0.99	0.99

Tabel 3 menunjukkan penurunan akurasi sebesar 0.18% (-0.0018) setelah *tuning*, namun perbedaannya tidak signifikan. Hasil pelatihan setelah tuning tetap menunjukkan akurasi 0.9977, yang mengindikasikan bahwa model sudah cukup optimal bahkan sebelum dilakukan penyesuaian *hyperparameter*. Kombinasi hyperparameter ini memastikan bahwa model mampu mengenali pola sentimen dalam tweet secara optimal, bahkan dalam kondisi dataset yang tidak seimbang [16]. Hasil ini menunjukkan bahwa model XGBoost-Vader mampu menangani ketidakseimbangan data dengan baik, sehingga *oversampling* tambahan tidak diperlukan. Optimasi *hyperparameter tuning* berhasil mengurangi *overfitting* dengan sedikit penurunan akurasi tanpa mengorbankan stabilitas dan performa model. Untuk memvisualisasikan dampak *tuning* terhadap performa model, dilakukan evaluasi menggunakan cross-validation.



Gambar 10. Cross Validation Accuracy per Fold Sebelum dan Setelah Tuning

Gambar 10 menunjukkan hasil cross-validation yang membandingkan performa model sebelum dan sesudah tuning. Akurasi tiap *fold* ditampilkan dalam dua garis berbeda: garis biru untuk model sebelum tuning dan garis merah putus-putus untuk model setelah tuning. Akurasi model sebelum tuning konsisten di angka 99,95% hingga 100%, sedangkan setelah tuning, akurasi awalnya sedikit lebih rendah tetapi kemudian stabil di angka 100%. Hasil ini menunjukkan bahwa model tetap memiliki stabilitas tinggi di seluruh *fold*, dengan variasi yang sangat kecil. Performa yang hampir sempurna ini mengindikasikan bahwa model mampu menggeneralisasi data dengan baik tanpa indikasi *overfitting* yang signifikan.

3.6 Analisis Hasil dan Implikasi

Hasil penelitian ini menunjukkan bahwa pendekatan *hybrid* XGBoost-VADER lebih efektif dalam menganalisis sentimen lowongan kerja di Twitter dibandingkan dengan metode sebelumnya. Integrasi model berbasis leksikon dengan *machine learning* terbukti mampu meningkatkan akurasi dalam analisis teks pendek [17]. Keunggulan utama penelitian ini terletak pada penerapan XGBoost-VADER sebagai metode analisis sentimen yang mengombinasikan keunggulan pendekatan leksikon dan *machine learning* untuk menghasilkan prediksi yang lebih akurat. Selain itu, teknik SMOTE Borderline diterapkan sebagai strategi penyeimbangan data

p-ISSN: 2775-4227
e-ISSN: 2775-4219

yang tidak hanya mengurangi risiko *overfitting*, tetapi juga meningkatkan stabilitas model dalam menghadapi skenario data yang tidak seimbang [7].

Penelitian sebelumnya oleh Pratama & Setyaningsih (2023) menggunakan Naïve Bayes dengan akurasi sentimen netral sebesar 64%, sementara Hidayat & Sugiyono (2023) membandingkan Naïve Bayes dan SVM, memperoleh akurasi 96,14% dan 94,80%, namun tanpa pendekatan hybrid. Asyaroh & Fitriani (2023) menerapkan Random Forest dan XGBoost dalam analisis kekerasan dalam rumah tangga, dengan XGBoost mencapai 86%, tetapi mengalami *overfitting*. Sementara itu, Widiarta et al. (2023) serta Yulistiani & Styawati (2024) menggunakan XGBoost dalam analisis kebijakan PPKM dan pemilu, menghasilkan akurasi 85,27% hingga 96%, namun tanpa mengombinasikan metode leksikon. Nurkarifin et al. (2024) meneliti sentimen pengguna aplikasi lowongan kerja menggunakan *Lexicon-Based Features* dan SVM, dengan akurasi berkisar 76% hingga 82%.

Dengan demikian, penelitian ini sejalan dengan berbagai studi terdahulu yang menerapkan metode *machine learning* dalam analisis sentimen. Namun, penelitian ini memberikan kontribusi lebih lanjut dengan pendekatan hybrid XGBoost-VADER, yang menggabungkan metode *lexicon-based* dan *machine learning* untuk meningkatkan performa analisis. Selain itu, penerapan SMOTE Borderline sebagai teknik penyeimbangan data menjadi aspek baru yang belum banyak digunakan dalam penelitian sebelumnya.

Meskipun tuning *hyperparameter* dalam penelitian ini menurunkan akurasi sebesar 0,18%, teknik ini berperan dalam mengurangi *overfitting* serta meningkatkan generalisasi model. Evaluasi menggunakan *classification report, confusion matrix*, dan *Stratified K-Fold Cross Validation* memastikan bahwa model yang dikembangkan memiliki konsistensi dan stabilitas terhadap berbagai distribusi data.

Untuk penelitian selanjutnya, disarankan eksplorasi teknik balancing data lain seperti ADASYN atau *Ensemble Resampling* guna membandingkan efektivitasnya dengan SMOTE Borderline. Selain itu, pengujian model dengan dataset lebih besar dan beragam dari berbagai platform media sosial diperlukan untuk menguji generalisasi model. Pendekatan *hybrid* yang lebih kompleks, seperti kombinasi XGBoost-VADER dengan model Transformer seperti BERT atau RoBERTa, juga dapat dieksplorasi untuk meningkatkan pemahaman konteks dalam analisis sentimen. Analisis aspek temporal dalam perubahan sentimen menjadi area potensial untuk memahami dinamika opini publik secara lebih mendalam.

4. KESIMPULAN

Penelitian ini mengusulkan pendekatan hybrid XGBoost-VADER untuk analisis sentimen lowongan kerja di Twitter. Kombinasi metode *lexicon-based* dan *machine learning* ini terbukti efektif dalam mengklasifikasikan sentimen positif, negatif, dan netral, terutama dalam menghadapi tantangan data yang tidak seimbang. Evaluasi model menunjukkan bahwa teknik SMOTE Borderline membantu menyeimbangkan distribusi kelas dalam dataset tanpa meningkatkan risiko *overfitting*. Namun, teknik ini tidak memberikan peningkatan akurasi yang signifikan, melainkan lebih berperan dalam memastikan model mampu mengenali kelas minoritas dengan lebih baik. *Hyperparameter* tuning menggunakan *GridSearchCV* menunjukkan adanya trade-off antara akurasi dan generalisasi model. Sebelum tuning, model mencapai akurasi 99.95%, sedangkan setelah tuning akurasi sedikit menurun menjadi 99.77%, dengan selisih 0.18%. Meskipun terjadi sedikit penurunan akurasi, tuning ini bertujuan untuk menghindari *overfitting* dan meningkatkan performa model pada data baru. Evaluasi menggunakan *confusion matrix, classification report,* dan *Stratified K-Fold Cross Validation* menunjukkan bahwa XGBoost-VADER tetap memiliki performa yang stabil dan akurat dalam klasifikasi sentimen lowongan kerja di Twitter.

Meskipun hasil penelitian ini menjanjikan, terdapat beberapa keterbatasan yang perlu diperhatikan. Salah satunya adalah keterbatasan dalam variasi dataset, yang dapat mempengaruhi generalisasi model terhadap data dari periode atau platform lain. Selain itu, meskipun pendekatan *hybrid* ini telah terbukti efektif, masih terdapat peluang eksplorasi lebih lanjut, seperti penerapan teknik *balancing* data lain seperti ADASYN atau *Ensemble Resampling*. Selain itu, penelitian selanjutnya dapat menguji model ini pada dataset yang lebih besar dan beragam, serta mengintegrasikan pendekatan berbasis Transformer seperti BERT atau RoBERTa untuk meningkatkan pemahaman konteks sentimen dalam teks pendek.

Dengan demikian, penelitian ini mengukuhkan XGBoost-VADER sebagai metode yang efektif dalam analisis sentimen lowongan kerja di Twitter, terutama dalam menangani karakteristik teks pendek dan variasi bahasa. Hasil penelitian ini dapat menjadi dasar untuk pengembangan lebih lanjut dalam analisis sentimen berbasis media sosial, khususnya dalam sistem rekrutmen berbasis AI atau pemantauan tren ketenagakerjaan secara real-time.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

DAFTAR PUSTAKA

[1] U. R. dan S. Yana, "Pengaruh Media Sosial Dalam Transformasi Pemasaran Digital," *JUPEIS: Jurnal Pendidikan dan Ilmu Sosial ,* vol. 3, no. 3, pp. 11-17, 2024

[2] D. I. A. Putri, F. T. Saputra dan R. Hardiyanti, "Pengaruh Penggunaan Media Sosial Twitter Terhadap Pemenuhan Kebutuhan," *Jurnal Ilmiah Wahana Pendidikan,* vol. 10, no. 8, pp. 410-418, 2024.

[3] 1 U. R. dan S. Yana, "Pengaruh Media Sosial Dalam Transformasi Pemasaran Digital," *JUPEIS: Jurnal Pendidikan dan Ilmu Sosial ,* vol. 3, no. 3, pp. 11-17, 2024.

[4] R. K. Pratama dan P. W. Setyaningsih, "Analisis Komentar Pada Twitter Terhadap Lapangan Kerja," *JURNAL INFORMATION SYSTEM & ARTIFICIAL INTELLIGENCE,* vol. 3, no. 2, pp. 217-227, 2023.

[5] A. Marcellino, Y. C. Moniung dan R. , "Penerapan Teknik SMOTE Pada Analisis Sentimen Bea," *Jurnal Algoritme,* vol. 4, no. 2, pp. 75-84, 2024.

[6] R. Asyaroh dan A. S. Fitriani, "Sentiment Analysis on Twitter About Domestic Violence Using Random," *KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen),* vol. 4, no. 4, pp. 1-9, 2023.

[7] R. E. H. Hermaliani dan M. Ernawati, "Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada," *Computer Science (CO-SCIENCE),* vol. 4, no. 1, pp. 80-88, 2024.

[8] W. Kurniawan dan U. Indahyanti, "Prediksi Angka Harapan Hidup Penduduk Menggunakan," *Indonesian Journal of Applied Technology,* vol. 1, no. 4, pp. 1-8, 20224.

[9] R. Aryanti, T. Misriati dan R. Hidayat, "Klasifikasi Risiko Kesehatan Ibu Hamil Menggunakan Random," *KLIK: Kajian Ilmiah Informatika dan Komputer,* vol. 3, no. 5, pp. 409-416, 2023.

[10] J. M. A. S. Dachi dan P. Sitompul, "Analisis Perbandingan Algoritma XGBoost dan Algoritma," *Jurnal Riset Rumpun Matematika dan Ilmu Pengetahuan Alam (JURRIMIPA),* vol. 2, no. 2, pp. 87-103, 2023.

[11] K. Akbar dan M. Hayaty, "Data Balancing untuk Mengatasi Imbalance Dataset pada," *Jurnal Ilmiah Intech : Information Technology Journal of UMUS,* vol. 2, no. 2, pp. 1-14, 2020.

[12] R. N. Ilyas, S. Mulyani, S. Wiguna dan M. Ramli, "Penggunaan Crawling Data X dengan Menggunakan Tweet," *TECHNOPEX-2024 Institut Teknologi Indonesia,* pp. 851-856, 2024.

[13] S. Khairunnisa, A. dan S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar," *JURNAL MEDIA INFORMATIKA BUDIDARMA,* vol. 5, no. 2, pp. 406-414, 2021.

[14] T. Ridwansyah, "Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia," *KLIK: Kajian Ilmiah Informatika dan Komputer,* vol. 2, no. 5, pp. 178-185, 2022.

[15] Z. A. Sriyanti, D. Y. K. Kartika dan A. R. E. Najaf, "Implementasi Model BERT Pada Analisis Sentimen Pengguna Twitter Terhadap Aksi Boikot Israel," *JITET (Jurnal Informatika dan Teknik Elektro Terapan),* vol. 12, no. 3, pp. 2335-2342, 2023.

[16] U. L. Yuhana, A. Purwarianti dan I. , "Tuning Hyperparameter pada Gradient Boosting untuk Klasifikasi Soal Cerita Otomatis," *JEPIN (Jurnal Edukasi dan Penelitian Informatika),* vol. 8, no. 1, pp. 134-139, 2022.

[17] B. Ramadhan dan F. S. Pane, "Pengaruh Hyperparameter Tuning untuk Efektivitas pada Pendekatan Hybrid dalam Mendiagnosis Stres dan Depresi: Tinjauan Studi Literatur," *Jurnal Tekno Insentif,* vol. 18, no. 2, pp. 104-118, 2024.