Volume 4 No. 1 (January - June 2024), pp. 01-18 | ISSN 2962-2360 (online) | ISSN 2962-3669 (print)
Arcitech: Journal of Computer Science and Artificial Intelligence
http://journal.iaincurup.ac.id/index.php/arcitech/article/view/10472 | DOI : http://dx.doi.org/10.29240/arcitech.v4i1.10472

Analisis Sentimen Opini Publik terhadap Kasus Korupsi Timah di Youtube Menggunakan Metode Oversampling dan Algoritma Decision Tree

Relin Pramudiya[1], Aldo Kadafi[2], Daniel Udjulawa[3]
[1,2,3]Program Studi Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang, Palembang, Indonesia
Email : relinrp@mhs.mdp.ac.id, aldokadafi@mhs.mdp.ac.id, daniel@mdp.ac.id

Article Information

Article history

Received 15 June 2024
Revised 23 June 2024
Accepted 30 June 2024
Available 30 June 2024

Keywords

Analisis Sentimen
Decision Tree
Korupsi
Opini publik
Smote Oversampling

Corresponding Author:

Relin Pramudiya,
Universitas Multi Data Palembang,
Email : relinrp@mhs.mdp.ac.id

Abstract

This study analyzes public opinion regarding the corruption case of PT. Timah (Tbk), which caused state losses of up to IDR 271 trillion, through YouTube comments. The methods used are the Decision Tree algorithm and the SMOTE Oversampling method. A total of 2501 comments were collected and processed. The stages include data preprocessing, sentiment labeling, and model training. The results show that the use of SMOTE improves the accuracy and performance of the model. With SMOTE, the model achieves an accuracy of 56%, a precision of 0.55, a recall of 0.55, and an F1-score of 0.55, while without SMOTE, the model only achieves 54%, a precision of 0.52, a recall of 0.52, and an F1-score of 0.52. Precision, recall, and F1-score also increase when using SMOTE. This study highlights the importance of the Oversampling technique in dealing with class imbalance to improve the accuracy and sentiment analysis model. These results make a significant contribution to sentiment analysis, highlighting the role of SMOTE in overcoming class imbalance and creating a more accurate model.

Keywords : *Corruption, Decision Tree, Public Opinion, Sentiment Analysis, SMOTE Oversampling*

Abstrak

Penelitian ini menganalisis opini masyarakat mengenai kasus korupsi PT. Timah (Tbk), yang menimbulkan kerugian negara hingga Rp 271 triliun, melalui komentar YouTube. Metode yang digunakan adalah algoritma Decision Tree dan metode Oversampling SMOTE. Sebanyak 2501 komentar dikumpulkan dan diolah. Tahapan meliputi preprocessing data, pelabelan sentimen, dan pelatihan model. Hasilnya menunjukkan bahwa penggunaan SMOTE meningkatkan akurasi dan performa model. Dengan SMOTE, model mencapai akurasi 56%, precision 0.55, recall 0.55, dan F1-score 0.55, sedangkan tanpa SMOTE, model hanya mencapai 54%, precision 0.52, recall 0.52, dan F1-score 0.52. Presisi, recall, dan skor F1 juga meningkat saat menggunakan SMOTE. Penelitian ini menyoroti pentingnya teknik Oversampling dalam menangani ketidakseimbangan kelas untuk meningkatkan akurasi dan model analisis sentimen. Hasil ini memberikan kontribusi yang signifikan terhadap analisis sentimen, menyoroti peran SMOTE dalam mengatasi ketidakseimbangan kelas dan menciptakan model yang lebih akurat.

Kata Kunci : *Analisis Sentimen, Decision Tree, Korupsi, Opini publik, SMOTE Oversampling*

1. Pendahuluan

Indonesia tengah dihebohkan oleh dugaan kasus korupsi dengan nilai mencapai ratusan triliun rupiah yang telah menggemparkan masyarakat. Industri pertambangan kali ini menjadi sorotan karena diduga terlibat dalam praktik tindak pidana korupsi. PT. Timah (Tbk) dituduh melakukan penambangan ilegal sejak tahun 2015 hingga 2022, yang mengakibatkan kerugian negara yang ditaksir mencapai Rp 271 triliun (Hanyfah et al., 2024). Dengan munculnya kasus ini banyak menimbulkan perbedaan pendapat dari masyarakat, karena kasus ini menimbulkan kerugian yang besar bagi negara, tentunya dari segi ekonomi, banyak komentar yang marah, dan ada pula yang bersikap netral terhadap kasus timah ini.

Kasus dugaan korupsi di PT. Timah (Tbk) yang melibatkan nilai mencapai ratusan triliun rupiah dan menimbulkan kerugian negara sebesar Rp 271 triliun merupakan isu yang sangat signifikan dan menggemparkan masyarakat Indonesia. Skandal ini tidak hanya berdampak besar pada perekonomian nasional, tetapi juga pada kepercayaan publik terhadap integritas industri pertambangan dan sistem hukum di Indonesia (Ramadhanti & Belitung, 2024). Oleh karena itu, penelitian mengenai sentimen masyarakat terhadap kasus ini sangat mendesak untuk memahami reaksi publik dan menyusun langkah-langkah penanganan yang tepat.

Dari banyaknya komentar opini masyarakat mengenai kasus ini, kami mengekstrak dan memanipulasi opini dari sekian banyak opini yang diungkapkan yang diambil dari video YouTube, banyaknya perbedaan pendapat membuat kasus ini semakin menarik untuk menciptakan sentimen berdasarkan opini tersebut. Banyak algoritma yang bisa digunakan untuk mengklasifikasi opini masyarakat menganai kasus timah ini. Salah satu algoritma tersebut adalah algoritma decision tree. Algoritma decision tree merupakan teknik pembelajaran mesin yang digunakan untuk pemodelan prediktif dan analisis data. Ini adalah algoritma pembelajaran berbasis pohon yang menghasilkan model yang dapat memprediksi nilai target berdasarkan serangkaian aturan keputusan yang sederhana (Singgalen, 2023).

Karena ada beberapa komentar yang tidak cocok antara negatif dan positif, maka kami menggunakan metode oversampling untuk meningkatkan jumlah komentar minoritas menjadi mayoritas agar pengelolaannya seimbang dan memiliki sentimen yang baik. *SMOTE* (Synthetic Minority Over-sampling Technique) merupakan salah satu teknik oversampling yang digunakan untuk menangani ketidakseimbangan kelas dalam dataset pada masalah klasifikasi. Data yang diambil ada 2501 komentar yang tidak seimbang antar positif dan negatif, untuk itu metode oversampling sebuah teknik yang digunakan untuk menangani masalah ketidakseimbangan kelas dalam dataset (Yudistira & Putra, 2021). Masalah ketidakseimbangan kelas terjadi ketika jumlah sampel dalam satu kelas jauh lebih sedikit dibandingkan dengan kelas lainnya dalam dataset. Ini dapat

menyebabkan model pembelajaran mesin menjadi bias terhadap kelas mayoritas dan kinerja model menjadi tidak optimal.

Penelitian sebelumnya terkait dengan sentimen masyarakat terhadap kasus korupsi di Indonesia mungkin sudah ada, namun belum ada yang secara spesifik memfokuskan pada kasus korupsi di PT. Timah (Tbk) dan bagaimana masyarakat merespon melalui komentar di platform media sosial seperti YouTube. Selain itu, penelitian mengenai penggunaan algoritma decision tree untuk klasifikasi sentimen komentar masyarakat dalam kasus korupsi pertambangan juga masih jarang ditemukan. Olah karena itu, penelitian ini akan berfokus pada analisis sentimen opini masyarakat terhadap kasus dugaan korupsi di PT. Timah (Tbk) dengan menggunakan metode oversampling dan metode decision tree.

2. Kajian Terdahulu

Penelitian yang dilakukan (Antrag et al., 2024) menjelaskan bahwa Penambangan timah ilegal di Bangka Buritung masih menjadi persoalan serius dan perlu ditanggulangi secara serius. Korupsi dalam administrasi pertambangan telah menyebabkan kerugian finansial yang besar bagi negara dan meningkatkan kerusakan lingkungan. Kajian ini bertujuan untuk mengkaji upaya penegakan hukum dan strategi penguatan sektor ekonomi non-pertambangan guna menciptakan tata kelola yang lebih transparan dan berkelanjutan. Metodologi penelitian yang digunakan didasarkan pada norma hukum dan mencakup analisis peraturan yang ada, studi kasus, dan wawancara dengan pemangku kepentingan terkait. Investigasi ini menunjukkan adanya kebutuhan mendesak untuk memperkuat institusi dan peraturan, termasuk pembentukan regulator independen yang berwenang mengawasi seluruh aspek penambangan timah dan menerapkan sanksi tegas atas pelanggaran yang ada. Studi ini menyimpulkan bahwa regulasi yang lebih kuat, diversifikasi ekonomi, dan pendekatan penegakan hukum yang menyeluruh dan komprehensif akan menghasilkan pengelolaan tambang timah yang lebih transparan, akuntabel, dan berkelanjutan, yang akan menekankan pada memberikan manfaat sebesar-besarnya kepada masyarakat dan lingkungan.

Penelitian yang dilakukan (A. Rahim et al., 2023) menjelaskan bahwa metode SMOTE dapat meningkatkan akurasi klasifikasi, penelitian ini menggunakan data pasien yang terdiagnosis penyakit jantung dan pasien tanpa penyakit jantung untuk meningkatkan klasifikasi penyakit jantung dengan menggabungkan Synthetic Minority Over-sampling Technique (SMOTE) dan algoritma random forest classifier. Hasilnya menunjukkan bahwa pendekatan ini dapat meningkatkan kemampuan model dalam mengidentifikasi kasus penyakit jantung. Evaluasi model menunjukkan peningkatan akurasi dibandingkan hasil penelitian sebelumnya, dengan akurasi sebesar 92%. Hasil terbaik ini lebih tinggi 2% dibandingkan hasil akurasi penelitian sebelumnya yang sebesar 90%.

Penelitian yang dilakukan (Harun & Putri Ananda, 2021) menjelaskan bahwa dari hasil analisis, Decision Tree digunakan untuk mengevaluasi akurasi dalam analisis data, terutama dalam tugas klasifikasi dan prediksi, dengan keunggulan dalam menghasilkan tingkat akurasi yang tinggi. Namun, dalam konteks analisis sentimen opini masyarakat terhadap vaksinasi COVID-19 di Indonesia, tampaknya ada lebih banyak tanggapan negatif daripada positif. Meskipun Naïve Bayes Classifier dan Decision Tree bisa digunakan untuk menganalisis sentimen dari komentar di Facebook, perbandingan akurasi keduanya menunjukkan perbedaan yang signifikan. Naïve Bayes Classifier mencapai akurasi 100.00%, sementara Decision Tree hanya mencapai 50.39%.

Korupsi dalam Industri Pertambangan

Korupsi adalah tindakan penyalahgunaan kekuasaan untuk keuntungan pribadi yang seringkali merugikan kepentingan public (Putri, 2021). Dalam industri pertambangan, korupsi dapat terjadi dalam berbagai bentuk, termasuk penambangan ilegal, penyuapan, dan manipulasi data produksi. Kasus korupsi yang melibatkan PT. Timah (Tbk) menuturkan, penambangan liar sudah berlangsung bertahun-tahun sehingga menimbulkan kerugian negara yang cukup besar. Korupsi di sektor ini tidak hanya menimbulkan kerugian ekonomi, namun juga merusak lingkungan dan menghambat pembangunan berkelanjutan.

Sentimen Analisis

Analisis sentimen adalah proses otomatis untuk menentukan apakah sebuah teks berisi opini yang positif, negatif, atau netral (Cahyaningtyas et al., 2021). Dalam kasus dugaan korupsi tersebut, PT. Timah (Tbk), analisis sentimen terhadap komentar masyarakat dapat memberikan gambaran mengenai opini masyarakat. Metode ini menggunakan algoritma pembelajaran mesin untuk mengklasifikasikan teks berdasarkan sentimen yang diungkapkan. Analisis ini penting karena dapat membantu memahami reaksi masyarakat terhadap isu tertentu dan mengidentifikasi aspek-aspek yang paling mempengaruhi opini publik.

Algoritma Decision Tree

Algoritma Decision Tree adalah teknik pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. Algoritma ini bekerja dengan cara membagi suatu kumpulan data menjadi subset-subset yang lebih kecil berdasarkan fitur-fitur tertentu, sehingga membentuk struktur pohon dengan node dan cabang. Setiap node mewakili fungsi atau atribut data, dan setiap cabang mewakili hasil keputusan. Algoritma ini populer karena interpretabilitasnya yang tinggi dan kemampuannya menangani data non-linier (Agus Trianto et al., 2023).

Ketidakseimbangan Kelas (Class Imbalance)

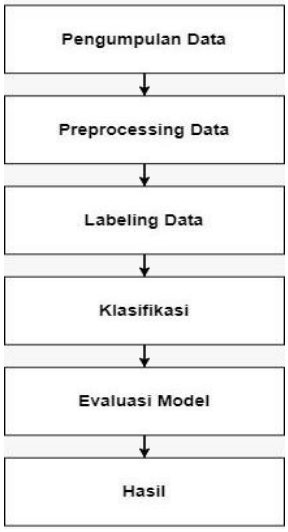
Ketidakseimbangan kelas terjadi ketika jumlah sampel dalam satu kelas jauh lebih sedikit dibandingkan kelas lain dalam kumpulan data. Hal ini dapat menyebabkan model pembelajaran mesin menjadi bias terhadap kelas mayoritas dan mengabaikan kelas minoritas, sehingga mengakibatkan performa model yang buruk dalam memprediksi kelas minoritas (Fadli et al., 2023). Dalam konteks analisis sentimen, ketidakseimbangan kelas dapat mempengaruhi akurasi dan keandalan hasil analisis.

SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE adalah teknik oversampling yang digunakan untuk memecahkan masalah ketidakseimbangan kelas dalam suatu dataset. Metode ini bekerja dengan membuat sampel sintetik dari kelas minoritas dengan melakukan interpolasi antar sampel yang ada. SMOTE meningkatkan jumlah sampel minoritas dengan membuat contoh baru yang serupa tetapi tidak identik, sehingga membantu meningkatkan performa model pembelajaran mesin (Pradana & Nooraeni, 2023). Dengan menggunakan SMOTE, model dapat dilatih pada kumpulan data yang lebih seimbang, sehingga meningkatkan akurasi prediksi dan mengurangi bias kelas mayoritas.

3. Metodologi Penelitian

Dalam penelitian ini, penulis membagi proses menjadi beberapa tahapan yakni:



Gambar 1. Tahapan Penelitian

3.1 Pengumpulan Data

Data tersebut dikumpulkan dari video channel YouTube Uya Kuya TV bertajuk "Kasus 271 T Timah, Sandra Dewi Bisa Jd Tersangka‼ Ada Aliran Dana Korupsi Ke Artis2 Besar Lain‼" yang membahas tentang kasus korupsi yang merugikan pendapatan negara. Komentar dikumpulkan menggunakan Netlytic

sebagai alat pengumpulan data (Kurniati et al., 2023). Netlytic digunakan untuk mengotomatisasi proses pengumpulan data dengan langkah-langkah sebagai berikut:

Pertama, video dipilih berdasarkan relevansinya dengan topik penelitian dan jumlah komentar yang signifikan. Kemudian, melalui Netlytic, autentikasi dilakukan dengan akun YouTube untuk memperoleh izin akses data komentar. URL video yang relevan dimasukkan ke dalam Netlytic, dan parameter pengumpulan data seperti periode waktu dan jumlah maksimal komentar yang diambil diatur sesuai kebutuhan penelitian. Netlytic secara otomatis mengumpulkan data komentar berdasarkan konfigurasi tersebut.

Dari proses ini, terkumpul sebanyak 2501 komentar. Data yang diperoleh mencakup berbagai atribut penting seperti id, author, description, guid, to, likecount, link, pubdate, replycount, title, dan authorChannelUrl. Atribut-atribut ini memberikan informasi detail tentang setiap komentar, seperti identifikasi unik, nama pengguna, isi komentar, jumlah 'like', tanggal dan waktu publikasi, jumlah balasan, dan URL kanal YouTube dari pengguna yang memberikan komentar. Dengan metode pengumpulan data yang terstruktur ini, data yang dihasilkan dapat digunakan untuk analisis sentimen yang lebih akurat dan dapat diandalkan.

3.2 Preprocessing Data

Pada tahap preprocessing, tahap pertama membersihkan data, seperti menghapus noise dan data yang tidak konsisten, serta transformasi data yang mengubah dan mengkonsolidasikan data ke dalam format yang sesuai. Tahap ini juga melibatkan reduksi data, termasuk seleksi dan ekstraksi fitur. Diharapkan setelah melalui tahap preprocessing, data telah menjadi final yang dianggap benar dan berguna untuk algoritma data mining (MZ et al., 2023). Data cleaning memiliki beberapa tahap seperti data normalization, stopword removal, tokenization, dan stemming.

Tahap pembersihan data (data cleaning) dilakukan dengan menghilangkan kolom-kolom yang tidak diperlukan dari kumpulan data untuk meningkatkan fokus dan kualitas data. Langkah ini juga menghapus data duplikat. Langkah selanjutnya adalah case folding yaitu mengubah seluruh huruf pada kolom deskripsi menjadi huruf kecil. Setelah itu dilakukan normalisasi untuk mengganti kata-kata yang tidak sesuai dengan ejaan bahasa Indonesia yang benar. Proses dilanjutkan dengan stopword removal, yaitu penghapusan kata-kata yang dianggap tidak penting terhadap topik, seperti “dan”, “di”, “ke”, “yang”, dan seterusnya. Tokenisasi kemudian dilakukan untuk membagi kalimat menjadi kata-kata. Terakhir, stemming dilakukan untuk menghilangkan imbuhan pada kata seperti “-an”, “-nya”, dan “me-”, guna mencari stem setiap kata dan menghilangkan karakter yang tidak diperlukan.

3.3 Labeling Data

Pada tahap ini berisi ulasan – ulasan yang telah diberi label sentimen yang bertujuan untuk mengklasifikasi ulasan tersebut mengarah kedalam kategori positif, negatif, atau netral. Proses Labeling data dilaksanakan secara manual dengan dibantu oleh seorang ahli bahasa sebanyak 2501 data. Setelah diberi label data disimpan dalam format excel diberi nama file KASUS_271_T_TIMAH (Permada & Sari, 2024).

3.4 Klasifikasi

Algoritma decision tree dimulai dengan memilih satu titik atau observasi dari data pelatihan yang diberi label. Decision tree kemudian mencari tetangga terdekat dari titik yang dipilih, yang jumlahnya telah ditentukan sebelumnya. Proses ini melibatkan penghitungan jarak antara titik yang dipilih dan tetangga terdekatnya menggunakan metrik jarak seperti jarak Euclidean atau jarak Manhattan. Label kelas yang paling sering muncul di antara tetangga terdekat akan digunakan sebagai label kelas untuk titik yang dipilih. Rumus Decision Tree terbagi menjadi 2 persamaan, persamaan (1) digunakan untuk mencari nilai Entropy dan persamaan (2) mencari nilai Gain (Pratiwi et al., 2024).

$$Entropy(S) = \sum n - pi * log\ pi \qquad (1)$$

Di mana pi menunjukkan persentase sampel kelas i dalam dataset D. Tingkat ketidakpastian atau ketidakteraturan set data diukur dengan entropi. Distribusi kelas semakin tidak teratur, semakin tinggi entropinya. Nilai tertinggi dari atribut yang tersedia digunakan untuk memilih atribut sebagai akar. Persamaan berikut ini menggunakan persamaan untuk menghitung Gain:

$$Gain\ (S) - \sum n\ |Si| * Entropy(SI) \qquad (2)$$

Pendekatan Decision Tree menggunakan $Entropy(Si)$ untuk menghitung kemampuan fitur untuk mempartisi data pada sebuah node. Sementara $\sum n|Si|$ merujuk pada penjumlahan untuk setiap subset dari Si, di mana $|Si|$ adalah jumlah sampel di setiap subset, entropi mengukur jumlah ketidakmurnian di node utama S.

Setelah melakukan preprocessing data, model decision tree dilatih menggunakan data pelatihan untuk menentukan nilai optimal dan membangun model klasifikasi. Setelah itu, data uji digunakan untuk mengevaluasi akurasi model. Dengan cara ini, model yang telah dibuat dapat digunakan untuk mengklasifikasikan data baru yang belum pernah dilihat sebelumnya.

3.5 Evaluasi Model

Confusion matrix digunakan untuk mengevaluasi kinerja algoritma klasifikasi dalam analisis sentimen, memberikan informasi apakah model memiliki performa yang baik atau tidak berdasarkan angka-angka yang dihasilkan. Confusion matrix terdiri dari empat komponen: True Positive (TP) adalah jumlah prediksi benar dengan nilai aktualnya juga benar; True Negative (TN) adalah jumlah prediksi benar dengan nilai aktualnya juga benar; False Positive (FP) adalah jumlah prediksi salah dengan nilai aktualnya salah; dan False Negative (FN) adalah jumlah prediksi salah dengan nilai aktualnya benar. Angka pada variabel TP dan TN merepresentasikan total prediksi benar oleh model, sedangkan angka pada variabel FP dan FN merepresentasikan total prediksi salah (Vanacore et al., 2024). Performa model dievaluasi dengan menghitung nilai accuracy, precision, recall, dan F1-Score. Akurasi mengukur sejauh mana model *klasifikasi* memberikan prediksi yang benar secara keseluruhan. Akurasi memberikan gambaran tentang seberapa baik model dapat mengklasifikasikan seluruh sampel dengan benar. Namun, dapat memberikan hasil yang bias jika proporsi kelas dalam *dataset* tidak seimbang. Rumus dari Akurasi dapat dihitung menggunakan Persamaan 3.6

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3.6)$$

Keterangan:

Tp = *True Positive*

Tn = *True Negative*

Fp = *False Positive*

Fn = *False Negative*

Precision mengukur sejauh mana prediksi positif yang dibuat oleh model adalah benar. Presisi penting ketika fokus pada mengurangi false positive. Ini memberikan informasi tentang seberapa baik model dapat mengidentifikasi kelas positif tanpa memberikan banyak kesalahan. Rumus Precisi dapat dilihat pada persamaan 3.7

$$Precission = \frac{TP}{TP+FP} \quad (3.7)$$

Keterangan:

Tp = *True Positive*

Fp = *False Positive*

3) Recall mengukur sejauh mana model dapat mendeteksi atau mengidentifikasi keseluruhan instance dari kelas positif. *Recall* penting ketika fokus pada mengurangi false negative. Ini memberikan informasi tentang seberapa baik model dapat mengenali seluruh *instance* yang seharusnya termasuk dalam kelas positif. Rumus Recall dapat dilihat pada persamaan 3.8

$$Recall = \frac{TP}{TP+FN} \qquad (3.8)$$

Keterangan:

Tp = True Positive

Fn = False Negative

4) *F1- Score* adalah matrik evaluasi yang merangkum performa model *klasifikasi* dengan mempertimbangkan kedua aspek penting, yaitu presisi (*Precision)* dan *recall. F1-Score* dirancang untuk memberikan gambaran holistik tentang kemampuan model dalam menangani kelas positif dan negatif. Rumus *F1-score* dapat dilihat pada (3.9)

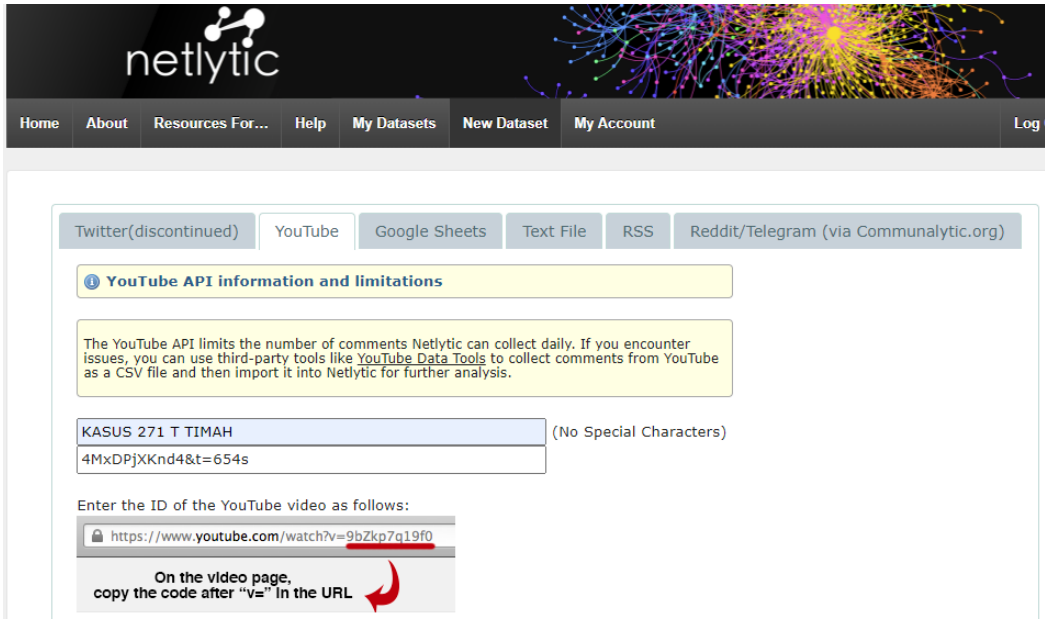
$$F1 - Score = \frac{2\ x\ Precission\ x\ Recall}{Precission + Recall} \qquad (3.9)$$

TP adalah banyaknya data dari kelas positif yang diprediksi secara tepat sebagai kelas positif.
FN adalah banyaknya data dari kelas positif yang salah diprediksi sebagai kelas negatif.
TN adalah banyaknya data dari kelas negatif yang diprediksi secara tepat sebagai kelas negatif.
FP adalah banyaknya data dari kelas negatif yang salah diprediksi sebagai kelas positif.

4. Hasil dan Pembahasan

4.1 Pengumpulan Data

Proses pengumpulan data menggunakan netlytic yang merupakan website tempat pengambilan data komentar dari youtube yang akan di gunakan sebagai dataset penelitian yang dapat dilihat pada gambar berikut:



Gambar 2. Proses pengumpulan data

4.2 Preprocessing Data

Pada tahap awal preprocessing data pada tahapan normalisasi, khususnya melalui fungsi "text_cleaning", diperoleh data sebagai berikut:

Tabel 1. Normalisasi Data

	nama_akun	text_cleaning	tanggal	sentimen
0	@erwinbinedy9772	rbt..dari jaman dahulu kala...susah untuk dit...	2024-05-06 0:25:29	positif
1	@deckypatria	coba bahas minyak	2024-05-05 5:06:44	netral
2	@user-rv3ol9ut3w	adaaaa banget	2024-05-04 15:20:18	netral
3	@IlhamSalbut	pemerintahan yang bobrok.	2024-05-04 10:18:46	negatif
4	@IlhamSalbut	lebih baik bubarkan indonesia dari pada rakyat...	2024-05-04 10:18:17	negatif

Pada tahap preprocessing data selanjutnya, khususnya melalui fungsi "text_cleaning" untuk menghilangkan stopword, diperoleh data sebagai berikut:

Tabel 2. Stopword Data

	nama_akun	text_cleaning	tanggal	sentimen
0	@erwinbinedy9772	rbt..dari jaman kala...susah telusuri nya..kar...	2024-05-06 0:25:29	positif

1	@deckypatria	coba bahas minyak	2024-05-05 5:06:44	netral
2	@user-rv3ol9ut3w	adaaaa banget	2024-05-04 15:20:18	netral
3	@IlhamSalbut	pemerintahan bobrok.	2024-05-04 10:18:46	negatif
4	@IlhamSalbut	lebih baik bubarkan indonesia rakyat slalu sen...	2024-05-04 10:18:17	negatif

Pada berikutnya dari preprocessing data, khususnya melalui fungsi "text_cleaning" untuk tokenisasi, diperoleh data sebagai berikut:

Tabel 3. Tokenisasi Data

	text_cleaning
0	[rbt..dari, jaman, kala...susah, telusuri, nya...
1	[coba, bahas, minyak]
2	[adaaaa, banget]
3	[pemerintahan, bobrok.]
4	[lebih, baik, bubarkan, indonesia, rakyat, sla...
	…
2495	[gk., peduli, mau, artis, mau, kolor, mlorot, ...
2496	[ngaku, mass, moeis, siapa, terlibat....,]
2497	[©, kalian, semua, gak, nyalahin, presiden,, ...
2498	[rusak, negoro, iki,]
2499	[beda, nya, alvin, lim, sama, ahok, beda, bang...

Pada tahap berikutnya dari preprocessing data, khususnya melalui fungsi "text_cleaning" untuk stemming, diperoleh data sebagai berikut:

Tabel 4. Stemming Data

	text_cleaning
0	rbt dari jaman kala susah telusur nya karena a...
1	cb bahas minyak
2	adaaaa banget
3	pemerintahan bobrok
4	lebih baik bubarkan indonesia rakyat sla...

Setelah melakukan preprocessing data melalui fungsi "text_cleaning" pada tahap sebelumnya, diperoleh data sebagai berikut:

Tabel 5. Hasil Preprocessing Data

No	Sebelum	Sesudah
1.	RBT..dari jaman dahulu kala...susah untuk di telusuri nya..karena di atas RBT..masih ada lg people power nya...	rbt dari jaman kala susah telusur nya karena atas rbt masih lagi people power nya
2.	pemerintahan yg bobrok.	perintah yang bobrok

3.	lebih baik bubarkan indonesia dr pada rakyat slalu di sengsarakan. dr dulu korup2 trus. di tangkap nyengar nyengir	lebih baik bubar indonesia dari rakyat selalu sengsara dari dulu korup2 trus tangkap nyengar nyengir
4.	Semoga Koh Alvin Lim Beserta keluarga Selalu Sehat, Dan Panjang Umurnya. Semoga Allah Memberkati Koh Alvin Lim Dan Keluarganya. Amin.!	moga koh alvin lim serta keluarga selalu sehat panjang umur moga allah kati koh alvin lim keluarga amin
5.	Mantap bang alvin	mantap bang alvin

4.3 Labeling Data

Proses pemberian label sentimen pada data yang telah diproses sebelumnya adalah sebagai berikut:

Tebel 6. Data yang sudah dilebeli

	nama_akun	text_cleaning	tanggal	sentimen
0	@erwinbinedy9772	rbt..dari jaman kala...susah telusuri nya..kar...	2024-05-06 0:25:29	positif
3	@IlhamSalbut	pemerintahan yg bobrok. Pemerintahan Yg Bobrok.	2024-05-04 10:18:46	negatif
4	@IlhamSalbut	lebih baik bubarkan indonesia dr rakyat slalu ...	2024-05-04 10:18:17	negatif
5	@partysarju7215	semoga koh alvin lim beserta keluarga selalu s...	2024-05-04 5:06:48	positif
6	@revanajah9094	mantap bang alvin	2024-05-04 4:50:03	positif
…	…	…	…	…
2493	@Faqih735	laa istrinya ikut nikmati dong,,, muatahil gak...	2024-04-05 11:46:10	negatif
2494	@kjijrjr	inilah sosok yg bergelar pahlawan rakyat.sosok...	2024-04-05 11:45:55	positif
2496	@Faqih735	ngaku mass moeis siapa terlibat....,	2024-04-05 11:44:45	negatif
2497	@user-ei8io8kn6k	kalian semua gak nyalahin presiden, padahal...	2024-04-05 11:44:04	negatif
2499	@gatotkaca8155	beda nya alvin lim sama ahok beda banget ...al...	2024-04-05 11:37:56	positif

4.4 Klasifikasi dan Evaluasi

Preprocessing awal data, khususnya fungsi "text_cleaning", menghasilkan data berikut. Proses pemberian label emosi pada data yang telah diproses sebelumnya dilakukan dengan cermat. Hasilnya kemudian divisualisasikan sebagai awan kata untuk kata-kata positif, memberikan

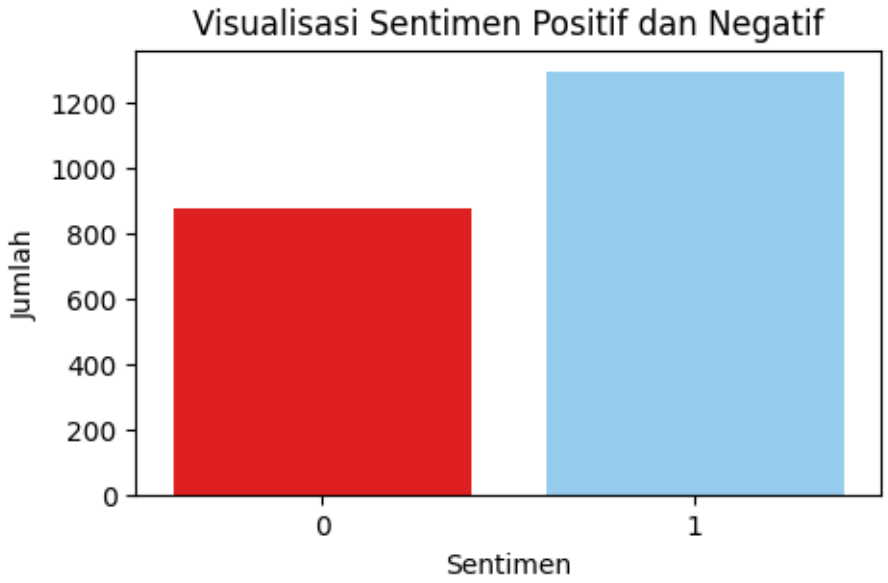
gambaran yang jelas tentang frekuensi dan pentingnya kata-kata tersebut dalam dataset.

Gambar 3. Visualisasi Kata Positif

Preprocessing pada data, khususnya pada fitur "text_cleaning", menghasilkan data tertentu. Proses pelabelan sentimen pada data yang sudah dipreproses dilakukan dengan teliti. Hasilnya kemudian divisualisasikan dalam bentuk wordcloud untuk kata-kata positif dan negatif, memberikan gambaran yang jelas mengenai frekuensi dan signifikansi kata-kata tersebut dalam dataset.

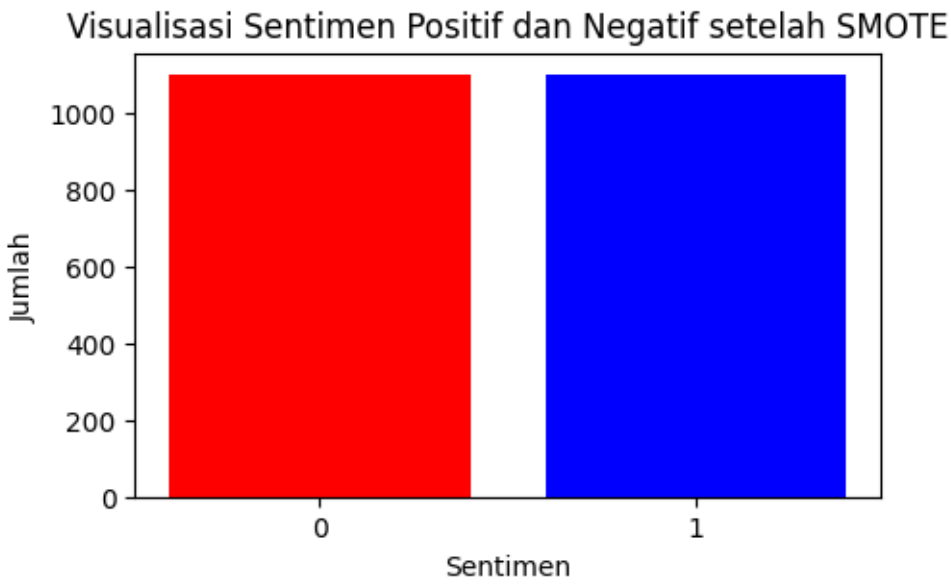
Gambar 4. Visualisasi Kata Negatif

Proses ini mencakup total 2166 data yang telah melalui tahap preprocessing, terdiri dari 873 data dengan sentimen negatif dan 1293 data dengan sentimen positif. Data tersebut kemudian digunakan sebagai bahan pelatihan untuk pengembangan model.



Gambar 5. Data awal yang digunakan untuk melatih model

Proses pengembangan model Decision Tree dimulai dengan menerapkan oversampling pada data menggunakan SMOTE untuk meningkatkan representasi kelas sentimen minoritas, memperbaiki ketidakseimbangan dalam dataset, dan meningkatkan kemampuan model dalam mengidentifikasi kelas yang kurang terwakili.



Gambar 6. Data hasil oversampling menggunakan SMOTE

Berikut disajikan table hasil accuracy, precision, recall, dan f1-score algoritma Decision Tree dan metode Oversampling menggunakan SMOTE dalam analisis sentimen opini publik terhadap kasus korupsi timah di youtube menggunakan metode Oversampling dan algoritma Decision Tree.

Tabel 7. Hasil dengan menggunakan SMOTE

	precision	recall	f1-score	support
negatif	0.47	0.51	0.49	134
positif	0.64	0.60	0.62	191
accuracy			0.56	325
macro avg	0.55	0.55	0.55	325
weighted avg	0.57	0.56	0.57	325

Hasilnya menunjukkan akurasi dengan menggunakan metode SMOTE dan algortima Decision Tree yang cukup tinggi, mencapai 0.56, yang menunjukkan kinerja yang baik dalam mengklasifikasikan data.

Tabel 8. Hasil tanpa menggunakan SMOTE

	precision	recall	f1-score	support
negatif	0.44	0.38	0.41	134
positif	0.60	0.66	0.63	191
accuracy			0.54	325
macro avg	0.52	0.52	0.52	325
weighted avg	0.54	0.54	0.54	325

Hasilnya menunjukkan akurasi tanpa menggunakan metode SMOTE dan hanya menggunakan algortima Decision Tree yang cukup tinggi, mencapai 0.54, yang menunjukkan kinerja yang baik dalam mengklasifikasikan data.

5. Kesimpulan

Penelitian ini menganalisis opini masyarakat mengenai kasus korupsi PT. Timah (Tbk) di YouTube menggunakan algoritma Decision Tree dengan metode oversampling SMOTE. Data yang dikumpulkan terdiri dari 2501 komentar yang tidak tercampur rata antara sentimen positif dan negatif. Setelah melalui serangkaian langkah preprocessing data, data tersebut sudah bersih dan siap digunakan.

Hasil klasifikasi menunjukkan bahwa penggunaan metode SMOTE meningkatkan akurasi dan performa model dalam mengklasifikasikan data sentimen. Dengan menggunakan metode SMOTE, algoritma Decision Tree mencapai akurasi 56%, precision 0.55, recall 0.55, dan F1-score 0.55. Sebaliknya, tanpa menggunakan metode SMOTE, algoritma Decision Tree mencapai akurasi 54%, precision 0.52, recall 0.52, dan F1-score 0.52. Hasil ini menunjukkan bahwa SMOTE secara efektif menangani ketidakseimbangan kelas dan meningkatkan performa model.

Secara keseluruhan, penelitian ini memberikan kontribusi penting terhadap analisis sentimen dengan menyoroti pentingnya penggunaan teknik oversampling seperti SMOTE untuk mengatasi ketidakseimbangan kelas dalam dataset. Dengan mengimplementasikan SMOTE, model yang dihasilkan menjadi lebih akurat dan kuat, sehingga memberikan pemahaman yang lebih baik tentang opini masyarakat terkait kasus korupsi PT. Timah (Tbk). Penelitian ini juga menggarisbawahi perlunya teknik preprocessing yang efektif dalam analisis sentimen untuk memastikan kualitas data yang tinggi dan hasil analisis yang lebih dapat diandalkan.

6. Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada semua pihak yang telah berkonstribusi pada penelitian ini.

7. Pernyataan Penulis

Penulis menyatakan bahwa tidak ada konflik kepentingan terkait publikasi artikel ini. Penulis menyatakan bahwa data dan makalah bebas dari plagiarisme serta penulis bertanggung jawab secara penuh atas keaslian artikel.

Bibliografi

A. Rahim, A. M., Inggrid Yanuar Risca Pratiwi, & Muhammad Ainul Fikri. (2023). Klasifikasi Penyakit Jantung Menggunakan Metode Synthetic Minority Over-Sampling Technique Dan Random Forest Clasifier. *Indonesian Journal of Computer Science*, *12*(5), 2995–3011. https://doi.org/10.33022/ijcs.v12i5.3413

Agus Trianto, G., Marzuki, M. F., Sihotang, T. Y., & Irsyad, H. (2023). *2 ND MDP Student Conference (MSC) 2023 Universitas Multi Data Palembang | 1 KLASIFIKASI* Opini Terhadap Resesi Indonesia 2023 Pada Twitter Menggunakan Algoritma Decesion Tree. 1–9.

Antrag, I. La, Situmaeng, Y. T., Arinda, S., & Rochim, A. A. (2024). *Penegakan Hukum Pertambangan Timah Ilegal Pasca Kasus Korupsi Tata Niaga Timah Di Bangka Belitung*. *3*(02), 184–191.

Cahyaningtyas, C., Nataliani, Y., & Widiasari, I. R. (2021). Analisis Sentimen Pada Rating Aplikasi Shopee Menggunakan Metode Decision Tree Berbasis SMOTE. *Aiti*, *18*(2), 173–184. https://doi.org/10.24246/aiti.v18i2.173-184

Fadli, M., Wijaya, V., Pribadi, M. R., & Widhiarso, W. (2023). Effect of TF-IDF Extraction and Application of SMOTE on Model Performance in Detecting Spam Email. *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, *November*, 637–641. https://doi.org/10.1109/EECSI59885.2023.10295851

Hanyfah, Z., Oktapia, A., & Tirta, M. (2024). Analisis Perhitungan Kerugian Negara dari Hasil Dugaan Tindak Pidana Korupsi yang dilakukan Oleh PT Timah (Tbk). *Journal of Law and Nation (JOLN)*, *3*(Mei), 351–358.

Harun, A., & Putri Ananda, D. (2021). Analisa Sentimen Opini Publik Tentang Vaksinasi Covid-19 di Indonesia Menggunakan Naïve bayes dan Decission Tree. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *1*(1), 58–64. https://doi.org/10.57152/malcom.v1i1.63

Kurniati, K., Kusmiati, H., & Rahmi, N. (2023). Analisis Hashtag UTBK-SNBT di Twitter Menggunakan Netlytic Tools. *Journal Computer Science and Information Systems : J-Cosys*, *3*(1), 44–48. https://doi.org/10.53514/jco.v3i1.383

MZ, Y., Boboring, J. E., & Fuadiah, N. (2023). Penerapan Metode K-Nearest Neighbor Dan Decision Tree Untuk Analisis Sentimen (Studi Kasus Mario Dandi). *Indonesian*

Journal Of Information Technology, 1–6.

Permada, D. N. R., & Sari, P. (2024). The effect of current ratio and debt to equity ratio on return on equity at PT. Timah Tbk. *Journal of Economics and Business Letters*, *4*(1), 43–53. https://doi.org/10.55942/jebl.v4i1.272

Pradana, R. S., & Nooraeni, R. (2023). Penerapan SMOTE pada Data Tidak Seimbang dalam Pemodelan Status NEET Penduduk Usia Muda di Provinsi Banten Tahun 2022. *Jurnal Kebijakan Pembangunan*, *18*(1), 91–104.

Pratiwi, S. A., Fauzi, A., Arum, S., Lestari, P., & Cahyana, Y. (2024). KLIK: Kajian Ilmiah Informatika dan Komputer Prediksi Persediaan Obat Pada Apotek Menggunakan Algoritma Decision Tree. *Media Online*, *4*(4), 2381–2388. https://doi.org/10.30865/klik.v4i4.1681

Putri, D. (2021). Korupsi Dan Prilaku Koruptif. *Jurnal Pendidikan, Agama Dan Sains*, *V*, 49–54.

Ramadhanti, I. R., & Belitung, U. B. (2024). *Jurnal Bevinding Vol 02 No 01 Tahun 2024 Fakultas Hukum Universitas Islam Batik Surakarta*. *02*(01), 28–43.

Singgalen, Y. A. (2023). Analisis Sentimen Top 10 Traveler Ranked Hotel di Kota Makassar Menggunakan Algoritma Decision Tree dan Support Vector Machine. *Media Online)*, *4*(1), 323–332. https://doi.org/10.30865/klik.v4i1.1153

Vanacore, A., Pellegrino, M. S., & Ciardiello, A. (2024). Fair evaluation of classifier predictive performance based on binary confusion matrix. *Computational Statistics*, *39*(1), 363–383. https://doi.org/10.1007/s00180-022-01301-9

Yudistira, N., & Putra, A. F. (2021). Algoritma Decision Tree Dan Smote Untuk Klasifikasi Serangan Jantung Miokarditis Yang Imbalance. *Jurnal Litbang Edusaintech*, *2*(2), 112–122. https://doi.org/10.51402/jle.v2i2.48