

EXTRACTING INSIGHTS FROM PLATFORM X (TWITTER): APPLICATION OF CRISP-DM AND TABLEAU VISUALIZATION IN SENTIMENT ANALYSIS OF THE FIRST PRESIDENTIAL AND VICE PRESIDENTIAL CANDIDATE DEBATE IN INDONESIA 2024

EKSTRAKSI WAWASAN DARI PLATFORM X (TWITTER): APLIKASI CRISP-DM DAN VISUALISASI TABLEAU DALAM ANALISIS SENTIMEN DEBAT CAPRES DAN CAWAPRES PERTAMA 2024 DI INDONESIA

Marastrika Farhan Nur Haiban¹, Nur Sultan Salahuddin^{2*}

Universitas Gunadarma^{1,2}

marastrikafarhan01@gmail.com¹, nur_sultan_s@yahoo.com²

ABSTRACT

Analysis of public sentiment regarding the first presidential-vice presidential debate in 2024 in Indonesia using data from Platform X (Twitter). This research uses the CRISP-DM methodology which includes the stages of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Data was collected using the Tweet-Harvest API crawling technique, then processed using Natural Language Processing (NLP) techniques to group sentiment into three categories, namely positive, negative and neutral. Modeling was carried out using Random Forest with a dataset division of 70% for training data and 30% for testing data. Even though the model shows high accuracy on the training data, there are indications of overfitting and oversampling on the testing data. To overcome this, undersampling and hyperparameter tuning techniques were applied which resulted in a model accuracy of 62%. The results of sentiment analysis are visualized using Tableau, resulting in an interactive dashboard that is easy to understand by technical and non-technical audiences. This Sentiment Analysis provides deep insight into the dynamics of public opinion during Indonesia's first presidential and vice presidential debate in 2024, with recommendations for further development to improve model performance and visualization quality.

Keywords: *Sentiment Analysis, CRISP-DM, Tableau, First Vice Presidential Candidate Debate, X (Twitter).*

ABSTRAK

Analisis sentimen publik terkait debat capres-cawapres pertama tahun 2024 di Indonesia menggunakan data dari Platform X (Twitter). Penelitian ini menggunakan metodologi CRISP-DM yang meliputi tahapan Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Data dikumpulkan menggunakan teknik crawling Tweet-Harvest API, kemudian diolah menggunakan teknik Natural Language Processing (NLP) untuk mengelompokkan sentimen ke dalam tiga kategori, yaitu positif, negatif, dan netral. Pemodelan dilakukan menggunakan Random Forest dengan pembagian dataset sebesar 70% untuk data training dan 30% untuk data testing. Meskipun model menunjukkan akurasi yang tinggi pada data training, namun terdapat indikasi overfitting dan oversampling pada data testing. Untuk mengatasi hal tersebut, diterapkan teknik undersampling dan hyperparameter tuning yang menghasilkan akurasi model sebesar 62%. Hasil analisis sentimen divisualisasikan menggunakan Tableau sehingga menghasilkan dashboard yang interaktif sehingga mudah dipahami oleh audiens teknis maupun non-teknis. Analisis Sentimen ini memberikan wawasan mendalam tentang dinamika opini publik selama debat presiden dan wakil presiden pertama di Indonesia pada tahun 2024, dengan rekomendasi untuk pengembangan lebih lanjut guna meningkatkan kinerja model dan kualitas visualisasi.

Kata Kunci: Analisis Sentimen, CRISP-DM, Tableau, Debat Capres Cawapres Pertama, X (Twitter)

PENDAHULUAN

Pada awal tahun 2024, Indonesia mencatat peningkatan penetrasi internet yang signifikan, dengan 221,6 juta dari total penduduk 278,7 juta jiwa, atau 79,5%, aktif menggunakan internet [1][2]. Hal ini menandai perluasan akses digital

yang substansial di berbagai wilayah, baik perkotaan maupun pedesaan. Fenomena ini terlihat dari semakin pentingnya peran media sosial dalam kehidupan sehari-hari, khususnya Platform X/Twitter, yang diperkirakan akan memiliki lebih dari 500 juta pengguna aktif bulanan pada tahun

2024 [3]. Platform ini tidak hanya berfungsi sebagai sarana interaksi sosial, tetapi juga sebagai alat penting dalam analisis sentimen publik, khususnya dalam konteks politik, seperti pada debat capres-cawapres pertama tahun 2024. Berdasarkan isu tersebut, penulis akan melakukan penelitian untuk mengetahui pandangan publik terhadap debat capres-cawapres pertama tahun 2024 di Indonesia menggunakan data dari Platform X/Twitter yang akan disajikan melalui visualisasi Tableau untuk informasi yang intuitif.

Penelitian sebelumnya tentang analisis sentimen dilakukan oleh [4] menggunakan data Twitter dengan metode berbasis leksikon dan random forest terkait prediksi hasil calon presiden. Hasil penelitian menunjukkan bahwa 97% publik Indonesia memiliki opini netral terhadap topik tersebut, dengan model random forest yang digunakan mencapai akurasi sebesar 88%. Penelitian lain yang dilakukan oleh [5] juga menggunakan data Twitter, menganalisis 714 tweet tentang Kecerdasan Buatan (AI) dengan pendekatan berbasis leksikon, menghasilkan sentimen positif sebesar 64%, negatif sebesar 25%, dan netral sebesar 11%. Penelitian analisis visualisasi yang dilakukan oleh [6] terhadap data prestasi mahasiswa tahun 2018-2020 di Universitas Lampung menunjukkan bahwa visualisasi data memberikan kontribusi positif, yaitu meningkatkan minat masyarakat terhadap Universitas Lampung. Selanjutnya, analisis sentimen dengan metode Berbasis Lexicon juga dilakukan pada tahun 2021 dengan 3108 data point menggunakan library Python twint tanpa API Twitter pada topik situs web pembajakan artikel penelitian, menghasilkan sentimen positif sebesar 59,51%, negatif sebesar 26,42%, dan netral sebesar 14,07%.

Kebaruan atau perbedaan pada penelitian ini terletak pada metodologi CRISP-DM yang meliputi tahapan Business Understanding, Data Understanding, Data Preparation,

Modeling, Evaluation, dan Deployment. Tujuannya adalah untuk memahami konteks debat capres dan cawapres pertama tahun 2024 dalam menilai sentimen publik mulai dari negatif hingga positif dan netral berdasarkan pasangan kandidat. Metode ini memadukan analisis sentimen berbasis leksikon dan model Random Forest, dengan hasil yang divisualisasikan menggunakan Tableau. Visualisasi ini dirancang untuk memudahkan pemahaman bagi pembaca tanpa latar belakang teknis, sehingga memungkinkan akses yang lebih intuitif dan komprehensif terhadap informasi tentang dinamika opini publik terkait debat pertama.

TINJAUAN PUSTAKA

Debat capres-cawapres merupakan komponen krusial dalam demokrasi elektoral, khususnya menjelang pemilihan umum. Di Indonesia, Komisi Pemilihan Umum (KPU) mengamanatkan debat ini sebagaimana diatur dalam Peraturan KPU Nomor 15 Tahun 2023. Debat ini bertujuan untuk memberikan wadah bagi para kandidat untuk menyampaikan visi, misi, dan program kerja mereka, sekaligus menguji kemampuan mereka dalam menanggapi isu-isu terkini yang kritis bagi masyarakat. Untuk pemilihan umum 2024, KPU akan menyelenggarakan lima debat dengan tema-tema khusus yang dirancang untuk menguji kompetensi dan tanggapan para kandidat terhadap berbagai isu, serta memberikan informasi penting bagi para pemilih untuk membuat keputusan yang tepat [7].

Dalam debat capres pertama, topik yang dibahas meliputi Tata Kelola Pemerintahan, Hukum, Hak Asasi Manusia, Pemberantasan Korupsi, Penguatan Demokrasi, Peningkatan Pelayanan Publik, dan Kerukunan Masyarakat. Fokus debat ini adalah untuk menilai pemahaman para kandidat tentang keadilan dan pemerintahan yang bersih, serta kemampuan mereka untuk mempromosikan kerukunan sosial. Debat-

debat ini tidak hanya menjadi ajang bagi para kandidat untuk menyampaikan visi dan misinya, tetapi juga untuk menguji kemampuan mereka dalam menanggapi isu-isu penting yang akan membentuk masa depan bangsa. Debat capres dan cawapres berperan penting dalam membantu para pemilih mengidentifikasi pemimpin yang paling sesuai dengan aspirasi mereka [8].

X/Twitter merupakan platform media sosial dan layanan mikroblog yang memungkinkan pengguna untuk mengunggah pesan singkat, yang disebut tweet, hingga 140 karakter secara real-time. Platform ini memungkinkan pengguna untuk berbagi pengalaman, mengikuti tren, informasi, dan berita global, serta tetap terhubung dengan orang lain. Setiap tweet bersifat publik dan dapat diakses oleh siapa saja, kapan saja, dan dari mana saja, serta secara otomatis muncul di linimasa pengikut. Istilah-istilah penting dalam penggunaan X/Twitter meliputi "Mention," untuk merujuk ke pengguna lain dengan '@'; "Hashtag," untuk menandai topik dengan '#'; "Emoticon," untuk mengekspresikan suasana hati dengan simbol; dan "Trending Topics," yang merujuk pada topik-topik diskusi yang sedang populer [9]. X/Twitter menyediakan API yang memudahkan peneliti untuk mengakses dan mengumpulkan data. API ini menawarkan berbagai endpoint untuk mencari tweet berdasarkan kata kunci, tagar, atau pengguna tertentu dan memungkinkan akses ke metadata seperti jumlah retweet, like, dan timestamp, yang berguna untuk analisis lebih lanjut [10].

Crawling merupakan metode pengumpulan konten informasi dari situs web, baik secara keseluruhan maupun bagian-bagian tertentu. Awalnya digunakan oleh browser web untuk mengisi indeks pencarian, tantangan dalam crawling meliputi ukuran data yang besar dan kurangnya kontrol terpusat atas konten [11]. Tweet Harvest atau Twitter Crawler merupakan alat baris perintah yang

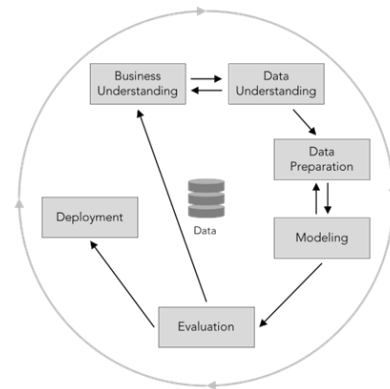
menggunakan Playwright untuk mengumpulkan data tweet dari Platform X/Twitter berdasarkan kata kunci dan rentang tanggal yang ditentukan. Tweet yang terkumpul disimpan dalam file CSV setelah menjalani autentikasi token API X/Twitter [12]. Data mining merupakan proses analisis yang bertujuan untuk mengeksplorasi data yang besar dan kompleks menggunakan metode yang terkait dengan pembelajaran mesin, statistik, dan sistem basis data. Fokus utamanya adalah mengidentifikasi pola yang tidak diketahui dari kumpulan data besar secara otomatis atau semi-otomatis, yang sering digunakan dalam bisnis untuk membuat keputusan strategis [13]. CRISP-DM (Cross-Industry Standard Process for Data Mining) adalah model penambangan data non-eksklusif, terdokumentasi, dan gratis yang dikembangkan melalui kolaborasi industri yang luas. Netral terhadap industri dan teknologi, model ini mendukung penerapan praktik terbaik dan menyediakan kerangka kerja untuk mengoptimalkan proyek penambangan data. CRISP-DM membagi prosesnya menjadi enam fase: Pemahaman Bisnis, Pemahaman Data, Persiapan Data, Pemodelan, Evaluasi, dan Penerapan, yang menawarkan panduan komprehensif untuk pelaksanaan proyek penambangan data, memastikan efektivitas dan efisiensi organisasi dalam mencapai hasil yang diinginkan [14]. Pemrosesan Bahasa Alami (NLP) adalah bidang dalam kecerdasan buatan yang berfokus pada interaksi antara komputer dan bahasa manusia. NLP bertujuan agar komputer memahami dan memproses bahasa manusia secara efektif. Tahapan dalam NLP meliputi pelipatan huruf besar-kecil, tokenisasi, penghilangan kata henti, stemming, dan penerjemahan, semuanya ditujukan untuk memproses teks atau ucapan sehingga dapat ditafsirkan oleh mesin. Proses ini penting untuk memungkinkan komputer memanipulasi bahasa alami untuk mendukung analisis

dan pemahaman yang berguna dalam konteks yang lebih luas [15].

Visualisasi data adalah teknik eksplorasi data yang mengubah data menjadi format visual, seperti grafik dan tabel, untuk memfasilitasi analisis dan penyajian. Teknik ini memungkinkan pemahaman data yang efektif dan kompleks serta menyederhanakan komunikasi informasi melalui representasi grafis yang intuitif. Visualisasi data sangat penting dalam analisis data kontemporer, membantu mengidentifikasi tren dan membuat data dapat diakses oleh pengguna dengan berbagai tingkat keahlian. Tableau adalah perangkat lunak visualisasi data yang populer dalam analisis data, yang memungkinkan pembuatan visualisasi interaktif dan informatif melalui antarmuka yang intuitif. Alat ini mendukung pemrosesan kumpulan data besar dan pembuatan grafik, peta, dan dasbor interaktif yang mendukung eksplorasi data dan pengambilan keputusan berbasis bukti [16]. Random Forest adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi dalam kategori pembelajaran terbimbing. Algoritma ini menggabungkan banyak pohon keputusan yang dibangun dari subset data acak untuk meningkatkan akurasi dan mengurangi risiko overfitting. Setiap pohon beroperasi secara independen, dan prediksi akhir dihasilkan melalui pemungutan suara dari semua pohon. Random Forest efektif dalam menangani kumpulan data yang tidak seimbang dan banyak fitur serta dapat memberikan estimasi pentingnya fitur. Pengaturan hiperparameter, seperti jumlah pohon ($n_estimator$), kedalaman maksimum (max_depth), dan jumlah fitur per pemisahan ($max_features$), sangat penting untuk kinerja optimal, biasanya dioptimalkan menggunakan GridSearchCV atau RandomizedSearchCV [17].

METODE

Penelitian ini mengkaji sentimen terkait debat capres dan cawapres pertama di Indonesia pada 12-21 Desember 2023 dengan menggunakan metodologi CRISP-DM. Metode ini mengorganisasikan proses analisis data ke dalam enam fase: Pemahaman Bisnis, Pemahaman Data, Persiapan Data, Pemodelan, Evaluasi, dan Penerapan. Setiap fase sangat penting untuk memastikan tujuan penelitian tercapai secara efektif. Analisis dilakukan dengan menggunakan Python, BigQuery, dan alat visualisasi Tableau. Penelitian ini memberikan gambaran terstruktur tentang penggunaan data dari Platform X (Twitter) untuk menganalisis dan memvisualisasikan sentimen selama periode debat, yang akan dijabarkan lebih lanjut pada Gambar 1.



Gambar 1. Diagram Alur CRISP-DM

Penelitian ini menerapkan metodologi CRISP-DM untuk menganalisis sentimen dari debat capres dan cawapres pertama tahun 2024 di Indonesia, menggunakan data yang diekstrak dari Platform X (Twitter). Prosesnya diawali dengan business understanding, yang berfokus pada identifikasi isu dan tujuan serta penerapan solusi efektif untuk memahami opini publik dan dampaknya terhadap persepsi masyarakat. Tahap data understanding dilakukan melalui teknik crawling menggunakan Tweet-Harvest API, dengan menggunakan kata kunci yang telah ditentukan sebelumnya untuk mengumpulkan tweet selama periode debat.

Fase penyiapan data meliputi pemrosesan dataset melalui pembersihan dan praproses data menggunakan teknik

NLP dan praproses dashboard untuk fase deployment. Pada fase pemodelan, data yang diolah digunakan untuk membangun model Random Forest, yang membagi data menjadi set pelatihan dan pengujian.

Selama fase evaluasi, model dievaluasi untuk mengidentifikasi oversampling dan overfitting menggunakan Random Search CV dan random under sampler. Fase terakhir, deployment, melibatkan pengintegrasian data ke dalam BigQuery dan pengembangan dashboard interaktif di Tableau untuk memvisualisasikan hasil analisis. Seluruh proses ini bertujuan untuk menghasilkan wawasan mendalam tentang dinamika opini publik selama debat presiden dan wakil presiden pertama tahun 2024 di Platform X (Twitter)

HASIL DAN PEMBAHASAN

Proses analisis data dalam penelitian ini mengikuti metodologi CRISP-DM, yang meliputi tahapan pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan penyebaran. Setiap fase bertujuan untuk mengatur dan mengoptimalkan ekstraksi dan analisis informasi dari data yang terkait dengan debat presiden dan wakil presiden pertama tahun 2024 di Platform X (Twitter), yang mengarah pada pengembangan pemahaman yang komprehensif dan aplikatif tentang sentimen publik.

A. Business Understanding

Pada tahap pemahaman bisnis, studi ini berfokus pada penguraian dinamika opini publik dan peningkatan efisiensi analisis data dari debat capres dan cawapres pertama tahun 2024 di Platform X (Twitter). Melalui penerapan analisis sentimen dan algoritma pemrosesan bahasa alami (NLP), penelitian ini mengidentifikasi dan mengklasifikasikan sentimen positif, negatif, dan netral, serta mengkategorikan kandidat berdasarkan sentimen tersebut. Tujuan utamanya adalah untuk memperdalam pemahaman tanggapan publik terhadap debat tersebut. Studi ini juga mengintegrasikan teknik

berbasis leksikon dan model hutan acak untuk memproses analisis data. Hasil analisis divisualisasikan menggunakan dasbor Tableau interaktif yang memfasilitasi eksplorasi data berdasarkan kandidat, lokasi, dan sentimen.

B. Data Understanding

Pada tahap pemahaman data, penelitian ini mengumpulkan data dari Platform X (Twitter) selama debat presiden dan wakil presiden pertama tahun 2024, dari tanggal 12 hingga 21 Desember 2023, menggunakan teknik crawling. Proses ini melibatkan integrasi Google Drive untuk penyimpanan data, yang diakses melalui Google Colab setelah mengimpor dan mengonfigurasi pustaka yang diperlukan seperti Pandas, NumPy, Seaborn, dan Matplotlib. Pustaka ini mendukung operasi data, analisis numerik, dan visualisasi. Data dikumpulkan menggunakan 12 kata kunci yang relevan dengan tema debat pertama, seperti 'debat pertama', 'debat presiden', 'debat wakil presiden', 'debat pemilu', 'pemilihan presiden 2024', 'tema tata kelola', 'tema hukum', 'tema hak asasi manusia', 'tema pemberantasan korupsi', 'tema penguatan demokrasi', dan 'tema kerukunan masyarakat'. Skrip Python yang disiapkan mengelola pengambilan tweet dengan pengaturan kueri dan batasan jumlah tweet yang dikumpulkan. Proses autentikasi dan pengambilan data yang diperlukan memastikan integritas dan kesiapan kumpulan data untuk analisis lebih lanjut, dengan penanganan kesalahan yang efektif untuk mengantisipasi dan memperbaiki potensi masalah data. Proses ini didokumentasikan dengan baik, dan hasilnya disajikan dalam Gambar 2 di bawah ini.

id_tweet	id_user	id_reply	id_retweet	id_quote_tweet	id_reply_tweet	id_retweet_tweet	id_quote_tweet_tweet
1	1775000000000000000	1775000000000000000	1775000000000000000	1775000000000000000	1775000000000000000	1775000000000000000	1775000000000000000
2	1775000000000000000	1775000000000000000	1775000000000000000	1775000000000000000	1775000000000000000	1775000000000000000	1775000000000000000
3	1775000000000000000	1775000000000000000	1775000000000000000	1775000000000000000	1775000000000000000	1775000000000000000	1775000000000000000
4	1775000000000000000	1775000000000000000	1775000000000000000	1775000000000000000	1775000000000000000	1775000000000000000	1775000000000000000

Gambar 2. Hasil Crawling Data

C. Data Preparation

Pada tahap persiapan data, data yang diperoleh dari perayapan debat presiden dan wakil presiden pertama menjalani pembersihan dataset, yang meliputi penghapusan data duplikat dan pemeriksaan nilai yang hilang dalam kerangka data. Selanjutnya, rekayasa fitur dilakukan dengan menambahkan kolom baru, memodifikasi konten variabel kolom, dan mengubah tipe data untuk analisis lebih lanjut, seperti yang diilustrasikan pada Gambar 3 di bawah ini.

Gambar 3. Data Cleaning

Langkah selanjutnya adalah tahap praproses NLP, yang melibatkan pembersihan data melalui proses seperti pembersihan teks, pelipatan huruf besar-kecil, dan tokenisasi, pemfilteran/penghapusan kata berhenti, stemming, penanganan kata tidak baku, kata baku, normalisasi, dan penerjemahan. Contoh hasil dari setiap proses praproses NLP dapat dilihat pada Tabel 1 di bawah ini.

Tabel 1. Contoh Hasil Praproses NLP

Teks	Tahapan Praproses NLP	Hasil Praproses NLP
@imanlagi Liat acara desak di ntb tadi di yutub sudah mulai kurang nendang. Pertanyaan nya hanya mengulang dari debat capres&ndesak Anies sebelumnya.	Cleaning Text	Liat acara desak di ntb tadi di yutub sudah mulai kurang nendang Pertanyaan nya hanya mengulang dari debat capres&ndesak Anies sebelumnya
	Case Folding dan Tokenization	['liat', 'acara', 'desak', 'di', 'ntb', 'tadi', 'di', 'yutub', 'sudah', 'mulai', 'kurang', 'nendang', 'pertanyaan', 'nya', 'hanya', 'mengulang', 'dari', 'debat', 'capres&ndesak', 'anies', 'sebelumnya']
	Stopwords Removal	['liat', 'acara', 'desak', 'ntb', 'yutub', 'nendang', 'nya', 'mengulang', 'debat', 'capres&ndesak', 'anies']

Stemming	liat acara desak ntb yutub nendang nya ulang debat capres&ndesak anies
Kata Tidak Baku	liat yutub
Kata Baku	lihat youtube
Normalisasi	lihat acara desak ntb youtube nendang nya ulang debat capres&ndesak anies
Translate	See the NTB Youtube urges to kick again the presidential debate urges Anies

Setelah menerapkan teknik pemrosesan bahasa alami (NLP), langkah selanjutnya adalah pelabelan sentimen berdasarkan leksikon menggunakan modul Vader. Data yang diproses dan diterjemahkan diklasifikasikan ke dalam tiga kategori sentimen: positif, negatif, dan netral. Klasifikasi ini didasarkan pada analisis sentimen dari tiga pasangan kandidat. Contoh hasil dari praproses leksikon ini disajikan dalam Tabel 2, yang menunjukkan distribusi sentimen untuk setiap kandidat.

Tabel 2. Contoh Hasil Praproses Berbasis Leksikon

Teks	Sentimen	Kandidat
In advance as a cagub anis signs the pact not to advance as a presidential candidate Prabowo Nyapres must just stay quiet	Netral	Paslon 1
Amin national team rises to electability aniesmuhammadin the presidential debate gives encouragement to prone to winning amen pairs of presidential elections	Positif	Paslon 1
After the first debate, the candidate for the name of Anies Baswedan Viral Social Media.Social Media Analysis Institution Emprit Drone Emprit Drone Sentiments Baswedan Negative Sentiments Winning Prabowo Subianto	Negatif	Paslon 2
I'm watching the vice presidential candidate debate when my sister's ganjar session dyes hair, sis, huh, it's not really good white.	Negatif	Paslon 3

Setelah fase praproses NLP dan leksikon, tahap praproses dasbor dilakukan. Tahap ini menggabungkan seluruh dataset yang telah dibersihkan dan dianalisis. Modifikasi seperti menambahkan kolom baru dan mengubah

nama serta isi kolom dilakukan untuk menyiapkan data untuk fase penyebaran, yang dapat dilihat pada Gambar 4.

id	sentiment	location	candidate	text
1	negatif	indonesia	prabowo	prabowo adalah orang yang tidak jujur
2	netral	indonesia	prabowo	prabowo adalah orang yang jujur
3	positif	indonesia	prabowo	prabowo adalah orang yang baik

Gambar 4. Hasil Data Preparation

D. Modelling

Dalam fase pemodelan, kumpulan data diklasifikasikan menggunakan model Random Forest, dengan pembagian 70% untuk data pelatihan dan 30% untuk pengujian. Meskipun model mencapai akurasi sempurna pada data pelatihan, efektivitasnya terbatas pada data pengujian dengan akurasi hanya 0,68, yang menunjukkan overfitting. Laporan klasifikasi menunjukkan bahwa model memiliki presisi tinggi tetapi recall rendah untuk kelas 'negatif', sering kali kehilangan kasus negatif. Sebaliknya, kelas 'positif' memiliki recall tinggi tetapi presisi lebih rendah, yang menunjukkan efektivitas dalam mengidentifikasi teks positif meskipun presisinya lebih rendah. Skor F1 yang berbeda antara kelas 'negatif' dan 'positif' menunjukkan ketidakseimbangan dalam manajemen kelas. Dominasi data kelas 'positif' dalam pengujian menunjukkan perlunya menyeimbangkan kembali distribusi kelas untuk meningkatkan efektivitas model. Kesimpulan dan analisis didokumentasikan dalam Gambar 5, yang menampilkan laporan klasifikasi, yang dengan jelas menunjukkan overfitting pada data pelatihan.

Random Forest Training Accuracy: 1.0
Random Forest Test Accuracy: 0.675876268862912

Random Forest Classification Report:

	precision	recall	f1-score	support
negatif	0.89	0.12	0.21	205
netral	0.62	0.69	0.65	209
positif	0.69	0.89	0.77	527
accuracy			0.68	941
macro avg	0.73	0.56	0.54	941
weighted avg	0.72	0.68	0.62	941

Gambar 5. Random Forest Data latih Classification Report

E. Evaluation

Pada fase pemodelan sebelumnya, oversampling dan overfitting pada model telah diidentifikasi. Untuk meningkatkan kinerja model Random Forest, penyetelan hiperparameter dilakukan menggunakan RandomSearchCV dan validasi silang, dan teknik pengambilan sampel diterapkan menggunakan random under sampler untuk mengatasi masalah ini. Kesimpulan dan analisis didokumentasikan dalam Gambar 6, yang menampilkan laporan klasifikasi, yang menunjukkan pengurangan overfitting pada model dan oversampling pada data.

Fitting 5 folds for each of 100 candidates, totalling 500 fits
Best parameters: {'n_estimators': 500, 'min_samples_split': 10, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': 40}
Best cross-validation score: 0.65
Test Accuracy: 0.6153018692879915

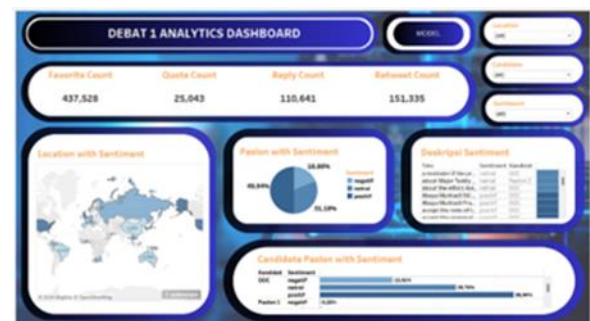
Classification Report:

	precision	recall	f1-score	support
negatif	0.47	0.59	0.52	205
netral	0.49	0.84	0.62	209
positif	0.87	0.54	0.66	527
accuracy			0.62	941
macro avg	0.61	0.65	0.60	941
weighted avg	0.70	0.62	0.62	941

Gambar 6. Random Forest Data Uji Classification Report

F. Deployment

Langkah selanjutnya adalah Visualisasi Data menggunakan Tableau untuk membuat dasbor yang menampilkan dinamika perdebatan dan kinerja model. Visualisasi ini menggunakan diagram tabel teks, diagram pai, diagram batang, diagram geografis, dan diagram pengukur, yang didokumentasikan dengan baik pada Gambar 7 dan 8.



Gambar 7. Dashboard Dinamika Debat Pertama

Gambar 7 menampilkan 'Dasbor Analisis Debat 1', yang dikembangkan menggunakan Tableau untuk menganalisis data dari debat presiden dan wakil presiden 2024. Dasbor ini menyajikan data

aktivitas media sosial terstruktur, termasuk 437.528 favorit, 25.043 kutipan, 110.641 balasan, dan 151.335 retweet. Peta interaktif menunjukkan distribusi geografis sentimen positif, netral, dan negatif. Diagram lingkaran menggambarkan distribusi sentimen di antara pasangan kandidat dengan 49,94% netral, 31,18% positif, dan 18,88% negatif. Tabel di dalam dasbor menyediakan contoh teks yang terkait dengan sentimen dan kandidat dan menampilkan persentase sentimen terperinci untuk setiap kandidat, yang menawarkan wawasan mendalam tentang reaksi publik. Fitur interaktif di dasbor, melalui navigasi 'Model', memungkinkan akses ke analisis model Random Forest, yang memperkaya pemahaman tentang dinamika sentimen dalam debat presiden dan wakil presiden pertama tahun 2024.



Gambar 8. Dashboard Model Random Forest

Gambar 8 menampilkan dasbor 'RANDOM FOREST CLASSIFIER' di Tableau, yang menyediakan analisis terperinci tentang kinerja model klasifikasi Random Forest. Judul di bagian atas menyoroti fokus pada klasifikasi data. Di sebelah kiri, tabel kinerja menampilkan metrik untuk tiga kelas: Negatif dengan skor F1 52%, presisi 47%, recall 59%; Netral dengan skor F1 62%, presisi 49%, recall 84%; dan Positif dengan skor F1 66%, presisi 87%, recall 54%. Tabel ini juga mencakup metrik Makro dan Tertimbang rata-rata. Di kanan atas, Bagan Gauge menunjukkan akurasi model pada 62%, menyediakan visualisasi kinerja model. Di bawah, kotak 'Insight' menyajikan analisis mendalam tentang kinerja model. Fitur navigasi 'OVERVIEW' memfasilitasi transisi

pengguna antara dasbor debat pertama dan dasbor model Random Forest, meningkatkan navigasi dan analisis yang lebih dalam

SIMPULAN

Studi ini menerapkan model Random Forest Classifier untuk menganalisis sentimen yang terkait dengan debat presiden dan wakil presiden pertama tahun 2024. Model ini dilatih menggunakan kumpulan data berbahasa Inggris yang dikumpulkan dari Twitter melalui API Tweet-harvest, dengan sentimen yang diklasifikasikan ke dalam tiga kategori: "Positif," "Negatif," dan "Netral." Sentimen ini juga dikaitkan dengan tiga pasangan kandidat yang diidentifikasi menggunakan kata kunci tertentu: "Paslon 1," "Paslon 2," dan "Paslon 3." Kumpulan data, yang terdiri dari 3.136 entri, dibersihkan dan dianalisis sebelum dibagi menjadi 70% untuk pelatihan dan 30% untuk pengujian.

Model Random Forest mengalami masalah dengan overfitting dan oversampling. Untuk mengatasi oversampling, peneliti menerapkan teknik undersampling untuk menyeimbangkan distribusi kelas, yang menghasilkan jumlah titik data yang sama, 484 untuk setiap kategori sentimen positif, negatif, dan netral. Dengan menggunakan Random Under Sampler dan penyetelan hiperparameter, model tersebut mencapai akurasi sebesar 62%. Hasil analisis sentimen dari debat pertama dan evaluasi kinerja model divisualisasikan menggunakan Tableau. Visualisasi ini dirancang untuk menyajikan informasi dengan cara yang mudah dipahami oleh audiens teknis dan non-teknis, memastikan bahwa temuan tersebut dapat diakses oleh semua pemangku kepentingan yang tertarik pada dinamika dan sentimen debat.

Meskipun penelitian ini menghasilkan temuan yang signifikan, penulis mengakui bahwa pengembangan lebih lanjut diperlukan untuk meningkatkan kinerja model dan kualitas

visualisasi dasbor. Secara khusus, meningkatkan akurasi model dalam mengklasifikasikan sentimen negatif dan netral sangatlah penting. Selain itu, disarankan untuk memasukkan wawasan yang lebih rinci ke dalam setiap visualisasi, yang memungkinkan pengguna untuk mendapatkan pemahaman yang lebih mendalam tentang data yang disajikan, lebih dari sekadar representasi visual.

DAFTAR PUSTAKA

- [1] Haryanto, A. T. (2024). APJII: Jumlah Pengguna Internet Indonesia Tembus 221 Juta Orang Baca artikel detikinet. Detikinet, February, 1–3.
- [2] Kominfo. (2024). Pengguna Internet Meningkat, Kominfo Galang Kolaborasi Tingkatkan Kualitas Layanan. Kominfo, 80.
- [3] Dean, B. (2024). X (Twitter) Statistics: How Many People Use X? (2024). Backlinko, November 2013. <https://backlinko.com/twitter-users>
- [4] Manullang, O., Prianto, C., & Harani, N. H. (2023). Analisis Sentimen Untuk Memprediksi Hasil Calon Pemilu Presiden Menggunakan Lexicon Based Dan Random Forest. *Jurnal Ilmiah Informatika*, 11(02), 159–169. <https://doi.org/10.33884/jif.v11i02.7987>
- [5] Putra, S. A., & Wijaya, A. (2023). Analisis Sentimen Artificial Intelligence (Ai) Pada Media Sosial Twitter Menggunakan Metode Lexicon Based. *JuSiTik: Jurnal Sistem Dan Teknologi Informasi Komunikasi*, 7(1), 21–28. <https://doi.org/10.32524/jusitik.v7i1.1042>
- [6] Al Malik, M. H., Andrian, R., & Utami, Y. T. (2023). Analisis Visualisasi Data Prestasi Mahasiswa Universitas Lampung Menggunakan Tableau. *Jurnal Pepadun*, 4(2), 158–165. <https://doi.org/10.23960/pepadun.v4i2.172>
- [7] Yonada Nancy. (2024). Rangkuman Hasil debat Capres-Cawapres 1-5, Siapa Pilihan Rakyat? Tirto.Id.
- [8] Mutiarasari, K. A. (2024). Tema dan Jadwal Debat Keempat Pilpres 2024, Cek di Sini. Detik.Com. <https://news.detik.com/pemilu/d-7134224/tema-dan-jadwal-debat-keempat-pilpres-2024-cek-di-sini>
- [9] Hadna, N. M. S., Paulus, I. S., & Winarno, W. (2016). Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen Di Twitter. *Seminar Nasional Teknologi Informasi Dan Komunikasi*, March, 1–8.
- [10] Riama santy sitorus, b. junedi hutagaol. (2023). Solusi DesainArsitekturSistem Informasi TerintegrasiDengan Pemanfaatan Media Sosial Untuk Strategi Peningkatan Kualitas Daya Tarik Wisata. *Journal of Tourism and Education*, 1–17.
- [11] Budiarto, J. (2024). Identifikasi Kebutuhan Masyarakat Nusa Tenggara Barat pada Pandemi Covid-19 di Media Sosial dengan Metode Crawling.pdf. *Jurnal Teknologi Informasi Dan Multimedia*.
- [12] Priansyah, E., & Sutabri, T. (2024). IJM: Indonesian Journal of Multidisciplinary Analisis Sentimen Berbasis Naïve Bayes Pada Media Sosial Twitter Terhadap Hasil Pemilu Indonesia 2024. *IJM: Indonesian Journal of Multidisciplinary*, 2, 128–138. <https://journal.csspublishing/index.php/ijm>
- [13] Papakyriakou, D., & Barbounakis, I. S. (2022). Data Mining Methods: A Review. *International Journal of Computer Applications*, 183(48), 5–19. <https://doi.org/10.5120/ijca2022921884>

- [14] Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, 181(2019), 526–534. <https://doi.org/10.1016/j.procs.2021.01.199>
- [15] Lauriola, I., Lavelli, A., & Aiolfi, F. (2022). An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools. *Neurocomputing*, 470(xxxx), 443–456. <https://doi.org/10.1016/j.neucom.2021.05.103>
- [16] Angreini, S. dan supratman edi. (2021). Visualisasi Data Lokasi Rawan Bencana Di Provinsi Sumatera Selatan Menggunakan Tableau | Jurnal Nasional Ilmu Komputer. Jurnal Nasional Ilmu Komputer. <http://journal.jis institute.org/index.php/jnik/article/view/528>
- [17] Kohli, S., & Joshi, P. (2021). “ A Brief Study on Random Forest Using Python .” *International Journal of Advances in Engineering and Management*, 3(6), 2063–2069. <https://doi.org/10.35629/5252-030620632069>