DOI: https://doi.org/10.37817/ikraith-informatika.v9i3
P-ISSN : 2580-4316
E-ISSN : 2654-8054

Analisis Sentimen Pengguna X Terhadap Pemilihan Gubenur Dki Jakarta Tahun 2024 Dengan Algoritma Naïve Bayes, K-Nearest Neighbor Dan Decision Tree

[1] Syarif Hidayatullah, [2] Arya Adhyaksa Waskita, [3] Achmad Hindasyah
[1] Teknik Informatika S-2, Universitas Pamulang, Tanggerang Selatan

E-mail: [1]haisyarif@gmail.com

ABSTRAK

Pemilihan Gubernur DKI Jakarta tahun 2024 merupakan salah satu agenda politik utama yang menarik perhatian publik Indonesia. Dalam konteks ini, analisis sentimen terhadap calon Gubernur dan Wakil Gubernur menjadi penting untuk memahami opini masyarakat. Dengan kemajuan teknologi, terutama internet dan media sosial seperti X (sebelumnya Twitter), masyarakat dapat dengan bebas mengungkapkan opini mengenai kandidat. Penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap ketiga pasangan calon Gubernur dan Wakil Gubernur DKI Jakarta melalui tweet yang dikumpulkan menggunakan metode crawling, menghasilkan 6.120 baris data, dengan masing-masing pasangan calon memperoleh 2.040 baris data. Perhitungan tingkat akusasi untuk sentiment analysis dilakukan menggunakan algoritma Naïve Bayes, K-Nearest Neighbor, dan Decision Tree untuk membandingkan akurasi ketiga algoritma tersebut dengan data testing – training 80:20 dan ekstraksi fitur menggunakan TF-IDF dan Transformer yang diimplementasikan menggunakan perangkat lunak RapidMiner. Berdasarkan hasil analisis, Naïve Bayes dengan representasi TF-IDF menunjukkan akurasi tertinggi untuk Paslon 1 sebesar 86,76%, diikuti oleh Paslon 2 dengan 76,96% dan Paslon 3 dengan 72,79%. Sementara itu, K-NN dengan TF-IDF memberikan hasil terbaik pada Paslon 1 (67,40%) dan Paslon 2 (71,32%), sedangkan Decision Tree memberikan akurasi tertinggi pada Paslon 1 dengan 72,55%. Untuk representasi Transformer, akurasi secara keseluruhan lebih rendah dibandingkan dengan TF-IDF, dengan Paslon 1 memperoleh akurasi 56,86%, Paslon 2 53,43%, dan Paslon 3 51,23%. Hasil ini menunjukkan bahwa TF-IDF lebih efektif dalam analisis sentimen terhadap tweet terkait calon Gubernur, dengan Naïve Bayes sebagai algoritma yang paling unggul dalam hal akurasi.

Kata kunci : Pemilihan Gubernur DKI Jakarta, Analisis Sentimen, Naïve Bayes, K-Nearest Neighbor, Decision Tree

ABSTRACT

The 2024 DKI Jakarta Gubernatorial Election is one of the main political agendas that attracts public attention in Indonesia. In this context, sentiment analysis of the Gubernatorial and Vice Gubernatorial candidates is important to understand public opinion. With the advancement of technology, especially the internet and social media platforms like X (formerly Twitter), the public can freely express their opinions regarding the candidates. This study aims to analyze public sentiment towards the three pairs of Gubernatorial and Vice Gubernatorial candidates of DKI Jakarta through tweets collected using the crawling method, resulting in 6,120 rows of data, with each candidate pair obtaining 2,040 rows of data. The accuracy of sentiment analysis was calculated using Naïve Bayes, K-Nearest Neighbor, and Decision Tree algorithms to compare the accuracy of these three algorithms with an 80:20 testing-training data split and feature extraction using TF-IDF and Transformer, implemented using the RapidMiner software. Based on the analysis results, Naïve Bayes with TF-IDF representation showed the highest accuracy for Paslon 1 at 86.76%, followed by Paslon 2 at 76.96%, and Paslon 3 at 72.79%. Meanwhile, K-NN with TF-IDF achieved the best results for Paslon 1 (67.40%) and Paslon 2 (71.32%), while Decision Tree achieved the highest accuracy for Paslon 1 at 72.55%. For the Transformer representation, the overall accuracy was lower compared to TF-IDF, with Paslon 1 achieving 56.86%, Paslon 2 at 53.43%, and Paslon 3 at 51.23%. These results indicate that TF-IDF is more effective for sentiment analysis of tweets related to the Gubernatorial candidates, with Naïve Bayes being the most accurate algorithm.

Keyword : DKI Jakarta Gubernatorial Election, Sentiment Analysis, Naïve Bayes, K-Nearest Neighbor, Decision Tree

P-ISSN : 2580-4316
E-ISSN : 2654-8054 DOI: https://doi.org/10.37817/ikraith-informatika.v9i3

1. PENDAHULUAN

Pastikan Anda menggunakan style Pemilihan Gubernur DKI Jakarta tahun 2024 menjadi salah satu agenda politik yang paling diperhatikan di Indonesia. Daerah Khusus Ibukota Jakarta akan menyelenggarakan pesta demokrasi yang sering dikenal sebagai Pemilihan Kepala Daerah (Pilkada). Masyarakat Jakarta akan memilih calon Gubernur dan Wakil Gubernur untuk memimpin Daerah Khusus Ibukota Jakarta selama lima tahun kedepan yaitu untuk periode 2024 – 2029. Masyarakat sering kali merasa bingung dalam menentukan pilihan berdasarkan kriteria apa yang harus diutamakan saat memilih calon Gubernur dan Wakil Gubernur. Oleh karena itu, menjelang Pemilihan Kepala Daerah (Pilkada) 2024 di DKI Jakarta, terdapat beberapa aspek yang perlu diperhatikan oleh masyarakat saat mengambil keputusan. Pilkada serentak akan dilaksanakan pada tanggal 27 November, melibatkan DKI Jakarta beserta 38 provinsi lainnya dan lebih dari 514 kabupaten/kota. Berdasarkan laporan dari Ketua KPU DKI Jakarta, Wahyu Dinata, yang disampaikan pada acara sosialisasi di Hotel Borobudur, Jakarta Pusat, pada tanggal 2 April 2024 (detikcom, 2024).

Kemajuan teknologi telah mempermudah masyarakat dalam mendapatkan sebuah informasi dengan sangat cepat, internet adalah sebuah bukti dari perkembangan teknologi yang saat ini banyak memberikan pengaruh terhadap masyarakat (Huda & Priyatna, 2019). Selama proses kampanye berlangsung dalam kurun waktu yang telah ditentukan, opini masyakarat terhadap setiap partai politik sangat berkembang secara masif melalui berbagai cara, salah satunya melalui media sosial. Berbagai opini masyarakat yang bersifat positif, negatif, ataupun netral dapat disampaikan secara bebas di media sosial (Adzhari et al., 2022). Pemilihan algoritma Naïve Bayes, K-Nearest Neighbor (KNN), dan Decision Tree dalam penelitian ini didasarkan pada karakteristik data tweet yang bersifat singkat, tidak terstruktur, serta mengandung variasi bahasa yang tinggi. Naïve Bayes digunakan karena memiliki kinerja yang baik dalam klasifikasi teks dan analisis sentimen dengan proses komputasi yang efisien, sehingga sesuai untuk pengolahan data dalam jumlah besar. Algoritma KNN dipilih karena mampu mengelompokkan data berdasarkan tingkat kemiripan antar tweet tanpa asumsi distribusi tertentu, sehingga efektif dalam menangkap pola opini yang serupa di media sosial. Sementara itu, Decision Tree digunakan karena mampu memodelkan hubungan yang kompleks antar fitur dan menghasilkan aturan keputusan yang mudah dipahami, sehingga dapat memberikan gambaran yang jelas mengenai faktor-faktor yang memengaruhi sentimen masyarakat.

2. LANDASAN TEORI

2.1 Pemilihan Kepala Daerah (Pilkada) DKI Jakarta 2024

Pemilihan Kepala Daerah baik tingkat provinsi maupun tingkat kabupaten dan kota dalam lingkup wilayah atau kawasan tertentu yang dilakukan secara serentak atau dalam waktu yang bersamaan.Pada Pilkada 2024 ini akan menjadi Pilkada serentak yang paling besar, dengan lebih dari 200 daerah yang akan melaksanakan pemilihan (Annisa, 2024), salah satu nya adalah DKI Jakarta.

KPU DKI Jakarta resmi menetapkan tiga pasangan calon untuk Pilkada Jakarta. Adapun ketiga paslon tersebut, yakni Ridwan Kamil (RK)-Suswono (RIDO), Dharma Pongrekun-Kun Wardana (Dharma-Kun), dan Pramono Anung-Rano Karno (Pram-Doel). Hal ini sesuai dengan keputusan KPU DKI No. 125 Tahun 2024 tentang penetapan pasangan calon peserta pemilihan Gubernur dan Wakil Gubernur DKI Jakarta 2024 (Putri & Dirgantara, 2024).

2.2 Data Mining

Secara sederhana data mining adalah penambahan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari

P-ISSN : 2580-4316
DOI: https://doi.org/10.37817/ikraith-informatika.v9i3 E-ISSN : 2654-8054

sejumlah data yang sangat benar. Data mining adalah suatu istilah yang digunakan untuk menemukan pengetahuan yang tersembunyi di dalam database (Pradnyana & Agustini, 2018). Istilah data mining memiliki beberapa padanan, seperti knowledge discovery ataupun pattern recognition. Kedua istilah tersebut sebenarnya memiliki ketepatannya masing-masing. Istilah knowledge discovery atau penemuan pengetahuan tepat digunakan karena tujuan utama dari data mining yaitu untuk mendapatkan pengetahuan yang tersembunyi di dalam bongkahan data (Arhami et al., 2020). Data mining dapat memproses seluruh data non-trivial untuk mengetahui pola dalam data, dimana yang ditemukan memiliki sifat sah atau mudah dipahami (Saputra & Primadasa, 2018). Knowledge Discovery and Data mining (KDD) adalah proses yang dibantu oleh komputer untuk menggali dan menganalisis sejumlah besar himpunan data dan mengekstrak informasi dan pengetahuan yang berguna. Data mining tools memperkirakan perilaku dan tren masa depan, memungkinkan bisnis untuk membuat keputusan yang proaktif dan berdasarkan pengetahuan. Data mining tools mampu menjawab permasalahan bisnis yang secara tradisional terlalu lama untuk diselesaikan. Data mining tools menjelajah database untuk mencari pola tersembunyi, menemukan infomasi yang prediktif yang mungkin dilewatkan para pakar karena berada di luar ekspektasi mereka (Saputra & Primadasa, 2018).

2.3 Naïve Bayes

Naïve Bayes merupakan sebuah pengklasifikasian probalistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma menggunakan teorema bayes dan mengansumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. Naïve Bayes juga didefinisikan sebagai pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Saleh, 2015).

Naïve Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Keuntungan penggunaan Naïve Bayes adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (Training Data) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. Naïve Bayes sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan (Saleh, 2015).

$$P(H \setminus X) = \frac{P(X \setminus H) \cdot P(H)}{P(X)}$$

Gambar 2.1 Rumus Persamaan NBC

Keterangan:
X : Data dengan class yang belum diketahui
H : Hipotesis data menggunakan suatu class spesifik
P(H|X) : Probabilitas hipotesis H berdasar kondisi X (parteriori probabilitas)
P(H) : Probabilitas hipotesis H (prior probabilitas)
P(X|H) : Probabilitas X bedasarkan kondisi pada hipotesis H
P(X) : Probabilitas H

2.4 K-Nearest Neighbor

K-Nearest Neighbor merupakan salah satu metode untuk mengambil keputusan menggunakan pembelajaran terawasi dimana hasil dari data masukkan yang baru diklasifikasi berdasarkan terdekat dalam data nilai. Algoritma K-Nearest Neighbor (KNN) adalah sebuah metode untuk melakukan kalsifikasi terhadap objek yang berdasarkan dari data pembelajaran yang jaraknya paling dekat dengan objek tersebut. KNN merupakan algotirma supervised learning dimana hasil dari query instance yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada algoritma KNN. Dalam K-Nearest Neighbor, kita mengelompokkan suatu data baru

P-ISSN : 2580-4316
E-ISSN : 2654-8054 DOI: https://doi.org/10.37817/ikraith-informatika.v9i3

berdasarkan jarak data baru ke beberapa data/tetangga (Neighbor) terdekat, dalam hal ini jumlah data/tetangga terdekat di tentukan oleh user yang dinyatakan dengan K-Nearest Neighbor merupakan salah satu model yang digunakan dalam algoritma supervised yang hasil instance diklasifikasikan berdasarkan mayoritas dari kategori pada K-Nearest Neighbor.

2.5 Decision Tree

Decisioin Tree merupakan pohon yang terstruktur dari kumpulin atribut untuk dapat d iuji dengan mempunyai tujuan memprediksi output-nya. Dimana setiap simpul internal menunjukkan tes pada atribut, hasil dari tes nya diwakili oleh masing-masing cabang, dan label kelas dipegang oleh setiap node-nya. Di dalam pohon keputusan node paling atas merupakan simpul akar. Menentukan akar pohon menggunakan gain tertinggi dari masing-masing atribut atau dengan menurut nilai index entrophy terendah.

$$\text{Entrophy(S)} = \sum_{i=0}^{n} -pi*log^2\ pi$$

$$\text{Gain(S,A)} = \text{Entrophy(S)} = \sum_{i}^{n} \frac{|si|}{|s|} * \text{Entrophy(Si)}$$

Decision Tree terbuat dari tiga simpul yaitu leaf, lalu terdiri juga dari simpul root yang merupakan titik awal dari suatu decision tree, dan yang terakhir adalah simpul perantara yang berhubungan dengan suatu pengujian (Puspita & Widodo, 2021).

2.6 Rapid Miner

RapidMiner adalah platform perangkat lunak data ilmu pengetahuan yang dikembangkan oleh perusahaan dengan nama yang sama, yang menyediakan lingkungan terpadu untuk pembelajaran mesin (machine learning), pembelajaran mendalam (deep learning), penambangan teks (text mining), dan analisis prediktif (predictive analytics). Aplikasi ini digunakan untuk aplikasi bisnis dan komersial serta untuk penelitian, pendidikan, pelatihan, pembuatan prototype dengan cepat, dan pengembangan aplikasi serta mendukung semua langkah proses pembelajaran mesin termasuk persiapan data, visualisasi hasil, validasi dan pengoptimalan. RapidMiner dikembangkan dengan model open core.

Menurut (CTI et al., 2017), RapidMiner merupakan software/perangkat lunak untuk pengolahan data. Dengan menggunakan prinsip dan algoritma data mining, RapidMiner mengekstrak pola-pola dari data set yang besar dengan mengkombinasikan metode statistika, kecerdasan buatan dan database. RapidMiner memudahkan penggunanya dalam melakukan perhitungan data yang sangat banyak dengan menggunakan operator-operator. Operator ini berfungsi untuk memodifikasi data. Data dihubungkan dengan node-node pada operator kemudian kita hanya tinggal menghubungkannya ke node hasil untuk melihat hasilnya. Hasil yang diperlihatkan RapidMiner pun dapat ditampilkan secara visual dengan grafik. Menjadikan RapidMiner adalah salah satu software pilihan untuk melakukan ekstraksi data dengan metode-metode data mining.

3. METODOLOGI

3.1 Crawling Data

Data dikumpulkan dari media sosial X menggunakan tools Tweet-Harvest. Data yang dikumpulkan berupa tweet yang mengandung opini masyarakat terhadap masing-masing pasangan Cagub-Cawagub DKI Jakarta. Tweet akan dikumpulkan berdasarkan kata kunci/keyword dan hashtag yang relevan dengan masing-masing pasangan Cagub-Cawagub tersebut.

Proses pengambilan data dari X dimulai dengan persiapan kata kunci/keyword yang akan di gunakan untuk mendapatkan tweet atau opini yang terdapat keyword tersebut. Setelah kata kunci/keyword telah di dikumpulkan, Langkah berikutnya adalah mendapatkan token autentikasi agar tools Tweet-Harvest dapat melakukan pencarian dan penyimpanan tweet atau opini sesuai dengan keyword yang dimasukkan.

Gambar 3.1 Token Autentikasi X

Setelah persiapan selesai, tahap berikutnya adalah memasukkan token autentikasi yang telah di dapatkan dan juga keyword yang akan

DOI: https://doi.org/10.37817/ikraith-informatika.v9i3
P-ISSN : 2580-4316
E-ISSN : 2654-8054

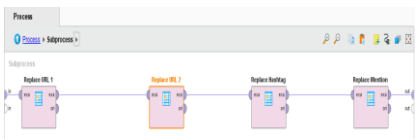
di cari. Tools Tweet-Harvest juga dapat menentukan kriteria pencarian seperti rentang tanggal kapan tweet dibuat, bahasa yang digunakan pada tweet, dan batas maksimum tweet yang akan diambil.

Gambar 3.2 Tools Twee-Harvest

Setelah kriteria pencarian ditetapkan, proses pengambilan data dijalankan dan data disimpan dalam format CSV. Data ini mencakup teks Tweet, metadata seperti waktu pembuatan dan ID Tweet, serta informasi pengguna seperti ID pengguna dan screen name. Untuk menangani batasan kuota data yang diambil, Tweet-Harvest dilengkapi script untuk otomatis menunggu dan melanjutkan kembali crawling data jika terkena rate limit pengambilan data setiap sekitar 10-15 menit (Theofilus Samosir, 2024).

3.2 Cleaning Data

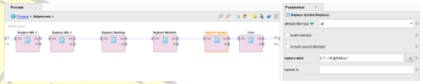
Cleaning atau pembersihan teks adalah tahap awal yang penting dalam preprocessing data tweet untuk memastikan bahwa data yang digunakan dalam analisis adalah relevan dan konsisten. Proses ini dimulai dengan menghapus elemen-elemen yang tidak diperlukan. Pada node Replace URL 1, digunakan untuk menghapus tautan (URLs) pada awal teks. Selanjutnya, node Replace URL 2, digunakan untuk menghapus tautan (URLs) pada tengah dan akhir teks. Setelah itu, digunakan node Replace Hashtag untuk menghilangkan hashtag (#) atau tanda pagar pada teks. Dan yang terakhir, digunakan node Replace Mention untuk menghilangkan mention (@username) pada teks. Hal ini dilakukan karena sering kali elemen-elemen di atas tidak memberikan informasi signifikan dalam analisis teks.



Gambar 3.3 Node Cleaning Data

3.3 Remove Symbol

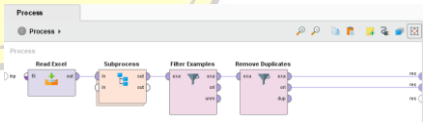
Selanjutnya, pembersihan teks melibatkan penghapusan karakter symbol, dan spasi ekstra, yang dapat memperbaiki format teks dan memudahkan proses analisis. Digunakan node Replace Symbol untuk menghilangkan simbol-simbol atau tanda baca atau elemen-elemen yang tidak relevan ([!-_?,.;:$%^&*()={}|]. Setelah itu, digunakan node Trim untuk menghilangkan spasi berlebih pada teks dan memperbaiki struktur teks, data yang dihasilkan menjadi lebih bersih dan siap untuk di analisis lebih lanjut.



Gambar 3.4 Node Remove Symbol

3.4 Remove Duplicate

Proses ini dilakukan untuk menghapus baris data yang identik (duplikasi) dalam sebuah dataset. Pertama-tama, digunakan node Filter Examples untuk memilih baris pada data yang tidak kosong setelah dilakukan cleaning, Setelah itu, node Remove Duplicates digunakan untuk memastikan bahwa hanya baris unik yang dipertahankan, sehingga mengurangi noise dan potensi bias dalam analisis data guna memperbaiki format teks dan memudahkan proses analisis. Pembersihan teks yang efektif memastikan bahwa model analisis atau Machine Learning dapat bekerja dengan data yang lebih konsisten dan akurat, meningkatkan kualitas hasil akhir dari analisis sentimen atau tugas analisis teks lainnya.



Gambar 3.5 Node Remove Duplicates

3.5 Pembobotan

Pembobotan data sentiment dalam penelitian ini menggunakan model VADER (Valence Aware Dictionary and sEntiment Reasoner), sebuah model yang dirancang untuk mengidentifikasi sentimen dalam teks berbahasa Inggris, terutama yang berasal dari media sosial dan format teks lain. Alat ini menghasilkan skor yang mencakup sentimen

P-ISSN : 2580-4316
E-ISSN : 2654-8054 DOI: https://doi.org/10.37817/ikraith-informatika.v9i3

positif, negatif, netral, dan skor komposit, memberikan gambaran menyeluruh tentang sentimen dalam teks yang dianalisis. VADER menggunakan metode berbasis leksikon yang dikombinasikan dengan aturan heuristik untuk menganalisis sentimen teks.

Gambar 3.6 Penggunaan Vader

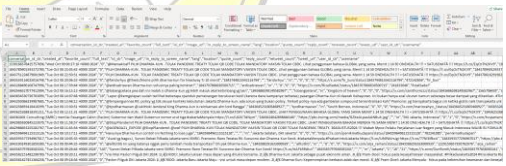
Pada proses nya, pertama-tama node Read CDV digunakan untuk membaca data file berformat csv yang akan di proses, setelah itu digunakan node Select Attributes untuk memilih kolom mana yang akan digunakan oleh VADER untuk di beri bobot. Selanjutnya, digunakan node Extract Sentiment untuk melakukan pembobotan menggunakan VADER, setiap teks dalam dataset dievaluasi menggunakan fungsi dari pustaka VADER. Alat ini menggunakan kamus yang memetakan kata-kata dan frasa umum ke nilai sentimen, serta memanfaatkan aturan gramatikal dan sintaksis yang spesifik untuk bahasa Inggris. Ketika digunakan untuk teks dalam bahasa Inggris, VADER mampu memberikan hasil analisis sentimen yang akurat dan andal. Namun, untuk teks dalam bahasa lain, hasilnya mungkin kurang akurat karena perbedaan dalam struktur bahasa dan konteks leksikal. Oleh karena itu, untuk mendapatkan hasil terbaik saat menggunakan VADER pada teks berbahasa non-Inggris, dalam penelitian ini teks akan diterjemahkan ke bahasa Inggris terlebih dahulu menggunakan Google Translate pada Google Spreadsheet seperti yang telah dijelaskan sebelumnya. Setelah dilakukan pembobotan menggunakan VADER, digunakan node Generate Attributes untuk menambah kolom baru dan merubah nilai hasil proses pembobotan menjadi 3 kategori, yaitu “Positive”, “Neutral”, dan “Negative”. Setelah hasil pembobotan di ubah menjadi sentiment, digunakan node Select Attributes kembali untuk memilih kolom pada data yang akan di proses selanjutnya, pada proses ini, kolom “Score” hasil pembobotan di hilangkan karena sudah di wakili oleh kolom “Sentiment”. Selanjutnya, digunakan node Nominal to Text untuk mengubah data yang berformat nominal ke format text. Lalu, untuk menyimpan data yang di prose, digunakan node Write Excel.

4. HASIL DAN PEMBAHASAN

4.1 Crawling Data

Saat melakukan crawling data menggunakan Tweet-Harvest, sangat penting untuk mempertimbangkan batasan penggunaan (rate limit) yang diberlakukan oleh X. X memiliki batasan jumlah permintaan yang dapat dilakukan dalam jangka waktu tertentu. Saat ini, X membatasi permintaan sebanyak 6000 tweet/hari untuk verified account, 600 tweet/hari untuk unverified account, dan 300 tweet/hari untuk pengguna baru. Untuk mengumpulkan tweet yang relevan, peneliti menggunakan kata kunci atau tagar tertentu dalam pada permintaan di Tweet-Harvest. Hasil dari proses crawling data yang tersimpan sebanyak 2.040 baris data tweet per masing-masing pasangan calon gubernur dan wakil gubernur DKI Jakarta. Proses crawling data dilakukan secara bertahap dikarenakan batasan yang diterapkan oleh X. Data yang di tampilkan pada screenshot setiap paslon dibawah hanya 20 data dari total masing-masing paslon 2.040 data yang di dapatkan

Gambar 4.1 Dataset Paslon 1



Gambar 4.2 Dataset Paslon 2

Gambar 4.3 Dataset Paslon 3

4.2 Cleaning Data

Pada bagian ini, dilakukan proses pembersihan teks yang bertujuan untuk menghilangkan elemen-elemen yang tidak

DOI: https://doi.org/10.37817/ikraith-informatika.v9i3
P-ISSN : 2580-4316
E-ISSN : 2654-8054

relevan dari data mentah yang diambil dari X.

Gambar 4.4 Data Sebelum Cleaning

Pertama-tama node Replace digunakan untuk menghapus tautan (URLs) di awal, di tengah dan di akhir tweet. Selanjutnya node Replace digunakan untuk menghilangkan hashtag (#) dan juga mention (@), memastikan hanya teks yang relevan yang tersisa.

Gambar 4.5 Hasil Cleaning Data

4.3 Remove Symbol

Setelah teks dibersihkan, selanjutnya adalah menghapus karakter symbol, dan spasi ekstra, yang dapat memperbaiki format teks agar memudahkan proses analisis.

Gambar 4.6 Data Sebelum Remove Symbol

Node Replace digunakan kembali untuk menghilangkan *symbol-symbol yang* tidak relevan dan memperbaiki struktur teks. Setelah itu digunakan *node* trim untuk menghilangkan spasi berlebih.

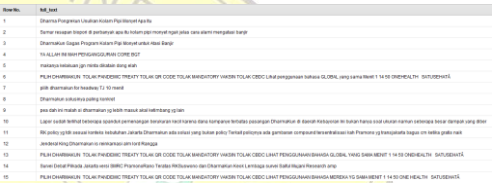
Gambar 4.7 Hasil Remove Symbol

4.4 Remove Duplicate

Proses ini dilakukan untuk menghapus data yang sama persis dalam dataset. Dengan begitu, hanya data yang unik yang disimpan, sehingga analisis menjadi lebih akurat dan mudah dilakukan. Hal ini juga membantu merapikan data agar lebih rapi dan siap untuk diproses lebih lanjut

Gambar 4.8 Data Sebelum Remove Duplicate

Juga node *Filter Example* digunakan untuk mengambil data yang tidak kosong/*blank*. Setelah semua *node* dijalankan, digunakan node *Write Excel* untuk nanti nya data yang telah di bersihkan di simpan dan di lakukan proses selanjutnya.



Gambar 4.9 Hasil Remove Duplicate

4.5 Translate

Pada proses ini, data dari masing-masing pasangan Calon Gubernur dan Wakil Gubernur di terjemahkan secara terpisah per masing-masing data. Penerjemahan dilakukan untuk menyiapkan data agar bisa dianalisis oleh model VADER, yang hanya bekerja dengan teks berbahasa Inggris. Menggunakan fitur Google Translate di Google Spreadsheet membuat proses ini lebih cepat dan mudah. Setelah diterjemahkan, data siap digunakan oleh VADER dan diberi bobot. Data yang di screenshot di bawah adalah 13 dari 2.040 data yang di proses translate.

Gambar 4.10 Hasil Transalte Data Paslon 1

P-ISSN : 2580-4316
E-ISSN : 2654-8054 DOI: https://doi.org/10.37817/ikraith-informatika.v9i3

Gambar 4. 1. Hasil *Translate* Data Paslon 2

Gambar 4. 2. Hasil *Translate* Data Paslon 3

4.6 Naïve Bayes

Hasil dari klasifikasi dataset Paslon 1 menggunakan algoritma Naïve Bayes, TF-IDF dan tools rapidminer dengan menggunakan hold-out sebagai pembagi data 80:20 memperoleh akurasi sebesar 86,76%. Adapun hasil pada tahap pengujian dapat dilihat pada tabel dibawah ini.

Tabel 4. 1. Hasil Pengujian Naïve Bayes hasil TF-IDF pada Dataset Paslon 1

	trueNeutral	True Negative	True Positive
pred. neutral	119	4	12
pred. negative	0	134	31
pred. positive	4	3	101

Dari tabel di atas dapat dijelaskan bahwa dataset Paslon 1 menggunakan metode Naïve Bayes hasil TF-IDF dengan perbandingan 80 data training dan 20 data testing maka menghasilkan prediksi sentimen Neutral pada kategori Neutral terdapat 119 data, untuk prediksi sentimen Negative pada kategori Negative terdapat 134 data, lalu untuk prediksi sentimen Positive pada kategori Positive terdapat 101 data.

Tabel 4. 2. Confussion Matrix Naïve Bayes hasil TF-IDF Dataset Paslon 1

Parameter	Nilai
Accuracy	86,76%
Recall	87,31%
Precision	87,63%
F1-Score	87,47%

Dari hasil pengujian metode Naïve Bayes hasil TF-IDF pada dataset Paslon 1 dengan perbandingan 80 data training dan 20 data testing maka mendapat hasil Akurasi 87,76% dari keseluruhan data, recall 87,31% dan precision 87,63% dan F1-Score 87,47%.

Untuk hasil dari klasifikasi dataset Paslon 2 menggunakan algoritma Naïve Bayes, TF-IDF dan tools rapidminer dengan menggunakan hold-out sebagai pembagi data 80:20 memperoleh akurasi sebesar 76,96%.

Tabel 4. 3. Hasil Pengujian Naïve Bayes & TF-IDF pada Dataset Paslon 2

	trueNeutral	True positive	True negative
pred. neutral	86	42	1
pred. positive	14	103	0
pred. negative	13	24	125

Dari tabel di atas dapat dijelaskan bahwa dataset Paslon 2 menggunakan metode Naïve Bayes hasil TF-IDF dengan perbandingan 80 data training dan 20 data testing maka menghasilkan prediksi sentimen Neutral pada kategori Neutral terdapat 86 data, untuk prediksi sentimen Positive pada kategori Positive terdapat 103 data, lalu untuk prediksi sentimen Negative pada kategori Negative terdapat 125 data.

Tabel 4. 4. Confussion Matrix Naïve Bayes & TF-IDF Dataset Paslon 2

Parameter	Nilai
Accuracy	76,96%
Recall	78,75%
Precision	77,29%
F1-Score	78,01%

Dari hasil pengujian metode Naïve Bayes hasil TF-IDF pada dataset Paslon 2 dengan perbandingan 80 data training dan 20 data

DOI: https://doi.org/10.37817/ikraith-informatika.v9i3
P-ISSN : 2580-4316
E-ISSN : 2654-8054

testing maka mendapat hasil Akurasi 76,96% dari keseluruhan data, recall 78,75% dan precision 77,29% dan F1-Score 78,01%.

Sementara itu, hasil dari klasifikasi dataset Paslon 3 menggunakan algoritma Naïve Bayes, TF-IDF dan tools rapidminer dengan menggunakan hold-out sebagai pembagi data 80:20 memperoleh akurasi sebesar 72,79%. Adapun hasil pada tahap pengujian dapat dilihat pada tabel dibawah ini.

Tabel 4. 5. Hasil Pengujian Naïve Bayes & TF-IDF pada Dataset Paslon 3

	True Positive	True negative	True neutral
pred. positive	94	6	27
pred. negative	31	126	15
pred. neutral	24	8	77

Dari tabel di atas dapat dijelaskan bahwa dataset Paslon 3 menggunakan metode Naïve Bayes hasil TF-IDF dengan perbandingan 80 data training dan 20 data testing maka menghasilkan prediksi sentimen Positive pada kategori Positive terdapat 94 data, untuk prediksi sentimen Negative pada kategori Negative terdapat 126 data, lalu untuk prediksi sentimen Neutral pada kategori Neutral terdapat 77 data.

Tabel 4. 6. Confussion Matrix Naïve Bayes & TF-IDF pada Dataset Paslon 3

Parameter	Nilai
Accuracy	72,79%
Recall	72,60%
Precision	72,64%
F1-Score	72,62%

Dari hasil pengujian metode Naïve Bayes hasil TF-IDF pada dataset Paslon 3 dengan perbandingan 80 data training dan 20 data testing maka mendapat hasil Akurasi 72,79% dari keseluruhan data, recall 72,60% dan precision 72,64% dan F1-Score 72,62%.

Hasil dari klasifikasi dataset Paslon 1 menggunakan algoritma Naïve Bayes, Transformer dan tools rapidminer dengan menggunakan hold-out sebagai pembagi data 80:20 memperoleh akurasi sebesar 56,86%. Adapun hasil pada tahap pengujian dapat dilihat pada tabel dibawah ini.

Tabel 4. 7. Hasil Pengujian Naïve Bayes hasil Transformer pada Dataset Paslon 1

	trueNeutral	True negative	True positive
pred. neutral	70	36	27
pred. negative	13	66	21
pred. positive	40	39	96

Dari tabel di atas dapat dijelaskan bahwa dataset Paslon 1 menggunakan metode Naïve Bayes hasil Transformer dengan perbandingan 80 data training dan 20 data testing maka menghasilkan prediksi sentimen Neutral pada kategori Neutral terdapat 70 data, untuk prediksi sentimen Negative pada kategori Negative terdapat 36 data, lalu untuk prediksi sentimen Positive pada kategori Positive terdapat 96 data.

Tabel 4. 8. Confussion Matrix Naïve Bayes hasil Transformer pada Dataset Paslon 1

Parameter	Nilai
Accuracy	56,86%
Recall	56,80%
Precision	57,83%
F1-Score	57,31%

Dari hasil pengujian metode Naïve Bayes hasil Transformer pada dataset Paslon 1 dengan perbandingan 80 data training dan 20 data testing maka mendapat hasil Akurasi 56,86% dari keseluruhan data, recall 56,80% dan precision 57,83% dan F1-Score 57,31%.

Untuk hasil dari klasifikasi dataset Paslon 2 menggunakan algoritma Naïve Bayes, Transformer dan tools rapidminer dengan menggunakan hold-out sebagai pembagi data 80:20 memperoleh akurasi sebesar 53,43%.

Tabel 4. 9. Hasil Pengujian Naïve Bayes hasil Transformer pada Dataset Paslon 2

	True neutral	True positive	True negative
pred. neutral	73	52	22

P-ISSN : 2580-4316
E-ISSN : 2654-8054 DOI: https://doi.org/10.37817/ikraith-informatika.v9i3

pred. positive	27	73	32
pred. negative	13	44	72

Dari tabel di atas dapat dijelaskan bahwa dataset Paslon 2 menggunakan metode Naïve Bayes hasil Transformer dengan perbandingan 80 data training dan 20 data testing maka menghasilkan prediksi sentimen Neutral pada kategori Neutral terdapat 73 data, untuk prediksi sentimen Positive pada kategori Positive terdapat 73 data, lalu untuk prediksi sentimen Negative pada kategori Negative terdapat 72 data.

Tabel 4. 10. Confussion Matrix Naïve Bayes hasil Transformer pada Dataset Paslon 2

Parameter	Nilai
Accuracy	53,43%
Recall	54,98%
Precision	53,59%
F1-Score	54,28%

Dari hasil pengujian metode Naïve Bayes hasil Transformer pada dataset Paslon 2 dengan perbandingan 80 data training dan 20 data testing maka mendapat hasil Akurasi 53,43% dari keseluruhan data, recall 54,98% dan precision 53,59% dan F1-Score 54,28%.

Sementara itu, hasil dari klasifikasi dataset Paslon 3 menggunakan algoritma Naïve Bayes, Tramsformer dan tools rapidminer dengan menggunakan hold-out sebagai pembagi data 80:20 memperoleh akurasi sebesar 51,23%. Adapun hasil pada tahap pengujian dapat dilihat pada tabel dibawah ini.

Tabel 4. 11. Hasil Pengujian Naïve Bayes hasil Transformer pada Dataset Paslon 3

	True positive	True negative	True neutral
pred. positive	83	53	30
pred. negative	43	68	31
pred. neutral	23	19	58

Dari tabel di atas dapat dijelaskan bahwa dataset Paslon 3 menggunakan metode Naïve Bayes hasil Transformer dengan perbandingan 80 data training dan 20 data testing maka menghasilkan prediksi sentimen Positive pada kategori Positive terdapat 83 data, untuk prediksi sentimen Negative pada kategori Negative terdapat 68 data, lalu untuk prediksi sentimen Neutral pada kategori Neutral terdapat 58 data.

Tabel 4. 12. Confussion Matrix Naïve Bayes hasil Transformer pada Dataset Paslon 3

Parameter	Nilai
Accuracy	51,23%
Recall	51,01%
Precision	51,96%
F1-Score	51,48%

Dari hasil pengujian metode Naïve Bayes hasil Transformer pada dataset Paslon 3 dengan perbandingan 80 data training dan 20 data testing maka mendapat hasil Akurasi 51,23% dari keseluruhan data, recall 51,01% dan precision 51,96% dan F1-Score 51,48%.

5. KESIMPULAN

Penelitian ini berhasil menganalisis sentimen opini publik terhadap pemilihan gubernur dan wakil gubernur DKI Jakarta tahun 2024 dengan menggunakan tiga algoritma klasifikasi Machine Learning, yaitu Naïve Bayes, K-Nearest Neighbors , dan Decision Tree. Data yang digunakan dalam penelitian ini diambil dari media sosial X, yang merupakan salah satu platform media sosial dengan volume data yang besar dan beragam. Proses analisis melibatkan tahap ekstraksi teks, preprocessing, dan pemodelan menggunakan algoritma-algoritma tersebut.
Penelitian ini mengungkapkan bahwa dari total 2.040 data tweet yang dianalisis untuk setiap pasangan calon, distribusi sentimen menunjukkan hasil yang bervariasi. Untuk paslon 1, sentimen positif mendominasi dengan 36,76%, diikuti oleh sentimen negatif sebesar 33,63%, dan sentimen netral sebanyak 29,61%. Sementara itu, pada paslon 2, sentimen positif mencapai 39,56%, diikuti oleh sentimen negatif sebesar 30,93%, dan sentimen netral sebanyak 29,51%. Sedangkan untuk paslon 3, sentimen positif tercatat sebesar 37,79%, sentimen negatif 32,94%, dan sentimen netral sebesar 29,26%.
Di antara algoritma yang diuji, algoritma Naïve Bayes dengan metode TF-IDF memberikan performa terbaik di antara semua kombinasi yang diuji, dengan akurasi tertinggi mencapai 86,76% pada Paslon 1. Naïve Bayes

secara konsisten unggul dibandingkan algoritma lainnya, terutama dalam menangani data yang diekstraksi menggunakan TF-IDF. Meskipun Decision Tree menunjukkan hasil yang lebih baik dengan Transformer, secara keseluruhan akurasinya masih di bawah Naïve Bayes dengan TF-IDF, menjadikan Naïve Bayes sebagai algoritma yang paling andal dalam menganalisis sentimen dibandingkan algoritma lainnya.

Penelitian ini memberikan kontribusi penting dalam memetakan sentimen publik terhadap pasangan calon dalam pemilihan calon gubernur dan calon wakil gubernur DKI Jakarta tahun 2024 melalui analisis data media sosial X, sekaligus menawarkan wawasan tentang efektivitas algoritma Machine Learning.

DAFTAR PUSTAKA

Adzhari, D. G., Sigit, A. A., Aditya, B. H., & Sari, E. P. S. P. (2022). Jejaring Sosial sebagai Wadah Pengenalan Wawasan Nusantara Era Digital di Kawasan Gunungpati Semarang. Jurnal Implementasi, 2(2), 128–136.

Akbar, Y., Nur Hannaa Regita, A., Sugiyono, S., & Wahyudi, T. (2024). Analisa Sentimen Pada Media Sosial " X " Pencarian Keyword ChatGPT Menggunakan Algoritma K-Nearest Abstrak. Jurnal Indonesia : Manajemen Informatika Dan Komunikasi (JIMIK), 5(3), 3291–3305.

Andreyestha, A., & Azizah, Q. N. (2022). Analisa Sentimen Kicauan Twitter Tokopedia Dengan Optimalisasi Data Tidak Seimbang Menggunakan Algoritma SMOTE. Infotek : Jurnal Informatika Dan Teknologi, 5(1), 108–116. https://doi.org/10.29408/jit.v5i1.4581

Annisa. (2024). Sejarah Pilkada Serentak di Indonesia dari Masa ke Masa. Https://Fahum.Umsu.Ac.Id/. https://fahum.umsu.ac.id/sejarah-pilkada-serentak-di-indonesia-dari-masa-ke-masa/

Anshor, A. H. (2022). Analisa Sentimen Warganet Terhadap KTT G20 Bali Menggunakan Algoritma Naïve Bayes. Jutisi : Jurnal Ilmiah Teknik Informatika Dan Sistem Informasi, 11(3), 819. https://doi.org/10.35889/jutisi.v11i3.1030

Aprilia, P. D., & Lestari, S. (2024). Analisa Sentimen Drama Korea Melalui Media Sosial X dengan Menggunakan Algoritma Naïve Bayes Abstrak. Jurnal Indonesia : Manajemen Informatika Dan Komunikasi (JIMIK), 5(3), 3248–3261.

Arhami, M., Kom, M., & Muhammad Nasir, S. T. (2020). Data Mining-Algoritma dan Implementasi. CV. Andi Offset.

Baba, U. R. H. (2024). Analisa Sentimen Menjelang Pemilihan Umum Presiden 2024 di Indonesia Menggunakan Perbandingan Performa Support Vector Machine (SVM) dan Naive Bayes. Innovative: Journal Of Social Science Research, 4(3), 11972–11990.

Fais Sya' bani, M. R., Enri, U., & Padilah, T. N. (2022). Analisis Sentimen Terhadap Bakal Calon Presiden 2024 Dengan Algoritme Naïve Bayes. JURIKOM (Jurnal Riset Komputer), 9(2), 265. https://doi.org/10.30865/jurikom.v9i2.3989

Hardi, N., Alkahfi, Y., Handayani, P., Gata, W., & Firdaus, M. R. (2021). Analisis Sentimen Physical Distancing pada Twitter Menggunakan Text Mining dengan Algoritma Naive Bayes Classifier. Sistemasi, 10(1), 131. https://doi.org/10.32520/stmsi.v10i1.1118

Hartanto, H. (2017). Text Mining dan Sentimen Analisis Twitter pada Gerakan LGBT. Intuisi: Jurnal Psikologi Ilmiah, 9(1), 18–25.

P-ISSN : 2580-4316
E-ISSN : 2654-8054 DOI: https://doi.org/10.37817/ikraith-informatika.v9i3

HATONANGAN, S. T., SUSANTO, A., & THOYYIBAH, T. (2024). Pengembangan Rancang Bangun Sistem Multi Control Penguncian Pintu Dengan Metode Prototype Berbasis Raspberry Pi (Studi Kasus: Tiara Kost). INFOTECH: JOURNAL OF TECHNOLOGY INFORMATION Учредители: Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Widuri, 10(2), 259-266.

Hidayat, E. Y., Hardiansyah, R. W., & Affandy, A. (2021). Analisis Sentimen Twitter untuk Menilai Opini Terhadap Perusahaan Publik Menggunakan Algoritma Deep Neural Network. Jurnal Nasional Teknologi Dan Sistem Informasi, 7(2), 108–118. https://doi.org/10.25077/teknosi.v7i2.2021.108-118

Khoiril Umat, Y. N., Nafsyi, D. R., Kusumaningsih, D., & Hakim, L. (2024). Analisis faktor yang mempengaruhi pemilihan gubernur daerah khusus jakarta menggunakan algoritma naive bayes dan regresi logistik 1). 9(2), 211–224.

Kusuma, A., & Nugroho, A. (2021). Analisa Sentimen Pada Twitter Terhadap Kenaikan Tarif Dasar Listrik Dengan Metode Naïve Bayes. Jurnal Ilmiah Teknologi Informasi Asia, 15(2), 137. https://doi.org/10.32815/jitika.v15i2.557

Lazuardi, M. T., Suprapti, T., & Wijaya, Y. A. (2023). PERANCANGAN MODEL SENTIMEN TWEET TERHADAP PILKADA DKI JAKARTA TAHUN 2017 MENGGUNAKAN ALGORITMA NAÏVE BAYES. 7(1), 308–312.

Maulana, F. A., & Ernawati, I. (2020). Analisa sentimen cyberbullying di jejaring sosial twitter dengan algoritma naïve bayes. Seminar Nasional Mahasiswa Ilmu Komputer Dan Aplikasinya (SENAMIKA, 529–538. https://conference.upnvj.ac.id/index.php/senamika/article/view/619

Muhammad Abdillah, Fikry, M., Yusra, Nazir, A., & Insani, F. (2024). Analisa sentimen terhadap kenaikan bbm di twitter (x) menggunakan naive bayes classifier. Jurnal CoSciTech (Computer Science and Information Technology), 5(1), 65–74. https://doi.org/10.37859/coscitech.v5i1.6954

Nadia, R., Muslim L, K., & Nhita, F. (2018). ANALISIS DAN IMPLEMENTASI ALGORITMA NAÏVE BAYES CLASSIFIER TERHADAPA PEMILIHAN GUBERNUR JAWA BARAT 2018 PADA MEDIA ONLINE. 5(1), 1678–1700.

Nardilasari, A. P., Hananto, A. L., Hilabi, S. S., Tukino, T., & Priyatna, B. (2023). Analisis Sentimen Calon Presiden 2024 Menggunakan Algoritma SVM Pada Media Sosial Twitter. JOINTECS (Journal of Information Technology and Computer Science), 8(1), 11. https://doi.org/10.31328/jointecs.v8i1.4265

Satria, H. (2024). Cara Mendapatkan Data (Crawl) Twitter X - Maret 2024. Https://Helmisatria.Com/Blog. https://helmisatria.com/blog/updated-crawl-data-twitter-x-maret-2024