Jurnal Ilmiah
Teknologi Sistem
Informasi

Jitsi

Vol. 6 (2025) No. 2

https://jurnal-itsi.org/index.php/jitsi; E-ISSN: 2722-4600; ISSN: 2722-4619
DOI: 10.62527/jitsi.6.2.332

Komparasi Tingkat Akurasi Sentimen Algoritma K-Nearest Neighbor Dan Naïve Bayes Pemilihan Gubernur Jawa Tengah 2024 di Sosial Media X

Fajar Ariyanto[#], Saefurrohman[#]

[#] *Fakultas Teknologi Informasi dan Industri, Universitas Stikubank Semarang, Jl. Tri Lomba Juang, Mugassari, Kec. Semarang Sel., Kota Semarang, Jawa Tengah 50241, Indonesia*
E-mail: kiyakiyo2017[at]gmail.com

A B S T R A C T S

This study highlights the importance of selecting appropriate algorithms for text data analysis and provides recommendations for future exploration of other machine learning and deep learning models to improve the accuracy of sentiment analysis. This research compares the accuracy level of the K-Nearest Neighbor (KNN) and Naïve Bayes algorithms in sentiment analysis in the 2024 Central Java gubernatorial election using data from the social media platform X (formerly Twitter). The data consists of 1,337 posts classified as positive or negative sentiment. Data crawling was done using RapidMiner, and analysis was done via Python in Google Colab. The research results show that the KNN algorithm achieves the highest accuracy of 81%, while Naïve Bayes has a maximum accuracy of 79%. The KNN algorithm is superior in handling text data because of the dependent calculations between attributes, while Naïve Bayes which uses independent calculations has slightly lower performance. This research provides insight into the reaction of public sentiment towards the candidate for governor of Central Java, where the Andhika-Hendi pair received more positive sentiment than Lutfi-Yasin.

Manuscript received Dec 29, 2024; revised Jun 17, 2025. accepted Jun 19, 2024 Date of publication Jun 30, 2025. International Journal, JITSI : Jurnal Ilmiah Teknologi Sistem Informasi licensed under a Creative Commons Attribution-Share Alike 4.0 International License

A B S T R A K

Studi ini menyoroti pentingnya memilih algoritma yang tepat untuk analisis data teks dan memberikan rekomendasi untuk eksplorasi model machine learning dan deep learning lainnya di masa depan guna meningkatkan akurasi analisis sentimen. Penelitian ini membandingkan tingkat akurasi algoritma K-Nearest Neighbor (KNN) dan Naïve Bayes dalam analisis sentimen pada pemilihan gubernur Jawa Tengah 2024 menggunakan data dari platform media sosial X (sebelumnya Twitter). Data terdiri dari 1.337 postingan yang diklasifikasikan sebagai sentimen positif atau negatif. Crawling data dilakukan menggunakan RapidMiner, dan analisis dilakukan melalui Python di Google Colab. Hasil penelitian menunjukkan bahwa algoritma KNN mencapai akurasi tertinggi sebesar 81%, sedangkan Naïve Bayes memiliki akurasi maksimal 79%. Algoritma KNN unggul dalam menangani data teks karena perhitungan dependen antar atributnya, sedangkan Naïve Bayes yang menggunakan perhitungan independen memiliki performa yang sedikit lebih rendah. Penelitian ini memberikan wawasan tentang reaksi sentimen masyarakat terhadap calon gubernur Jawa Tengah, di mana pasangan Andhika-Hendi menerima lebih banyak sentimen positif dibandingkan Lutfi-Yasin.

Keywords / Kata Kunci — ***Comparison, Sentiment Accuracy, KNN, Naïve Bayes***

CORRESPONDING AUTHOR

Fajar Ariyanto
Fakultas Teknologi Informasi dan Industri, Universitas Stikubank Semarang, Jl. Tri Lomba Juang, Mugassari, Kec. Semarang Sel., Kota Semarang, Jawa Tengah 50241, Indonesia
Email: buwasbanget[at]gmail.com

1. PENDAHULUAN

Elektabilitas dalam ranah politik sering kali menjadi topik hangat yang mendapat perhatian luas dari masyarakat, yang secara tidak langsung memengaruhi individu atau kelompok tertentu. Berita terbaru dan informasi yang mudah diakses dari berbagai sumber digital mendorong masyarakat untuk memberikan tanggapan, yang sering kali diterjemahkan sebagai umpan balik bagi tokoh-tokoh atau kelompok tertentu. Media sosial Twitter, yang kini dikenal sebagai X, sering kali menjadi pusat perbincangan tren terkait isu-isu skala nasional maupun internasional, dan menjadi wadah bagi netizen untuk menyuarakan opini mengenai berbagai isu yang tengah ramai dibahas di jaringan sosial yang sangat kompleks ini. Data menunjukkan bahwa pengguna X di Indonesia mencapai 24,69 juta dari total 600 juta pengguna global (https://www.cnbcindonesia.com/tech).

Sosial media X telah menjadi platform utama untuk berdiskusi dan berbagi opini tentang isu-isu politik. Memahami sentimen pengguna di media sosial bisa memberikan gambaran lebih jelas tentang opini publik. Sentimen pengguna media sosial dapat dianalisis menggunakan berbagai algoritma machine learning. Algoritma klasifikasi sudah banyak digunakan untuk membantu dalam pengolahan data seperti Random Forest, Decision Tree, Naïve Bayes, K-Means dan juga Support Vector Machine (SVM) (Novianti et al., 2022). Metode di atas yang diaplikasikan dalam penelitian ini adalah perbandingan keakurasian mengenai metode Naive Bayes dan K-Nearest Neighbor (KNN), pengklasifikasian Naive Bayes termasuk dalam keluarga pengklasifiksian probabilistik berdasarkan teori bayes. Fitur utama dari klasifikasi ini adalah asumsi bahwa semua variabel independen bersyarat yang menjadi alasan untuk menyebutkan naive. Parameter klasifikasi Naive Bayes dapat dipelajari secara terpisah, lebih sederhana dan lebih cepat (Ilić et al., 2022) dan memiliki keakuratan lebih baik dibanding metode data mining lainnya (Hozairi et al., 2021).

Pengklasifikasian K-Nearest Neighbor (KNN) adalah metode untuk mengklasifikasikan objek berdasarkan kedekatannya dengan objek lain dalam data pelatihan. Keakuratan algoritma KNN sangat bergantung pada keberadaan karakteristik yang tidak relevan atau ketika bobot fitur tidak mencerminkan relevansinya terhadap klasifikasi (Nuqoba & Djunaidy, 2014). Algoritma KNN memiliki beberapa keunggulan, di antaranya pelatihan yang sangat cepat, kesederhanaan dan kemudahan untuk dipahami, ketahanan terhadap data pelatihan yang tidak beraturan, serta efektivitas pada dataset yang besar. Namun, algoritma ini juga memiliki kelemahan, seperti bias nilai k, kompleksitas komputasi, keterbatasan memori, dan rentan terhadap atribut yang tidak relevan (Rosso, 2019). Penelitian terdahulu mengenai perbandingan algoritma data mining Naive Bayes dan KNN yaitu penelitian yang diteliti oleh Syarifuddinn (2020) yang menjelaskan tentang penyelarasan dan pandangan baru mengenai suatu isu dalam twitter yang memiliki kecenderungan opini masyarakat terhadap twitter condong positif dengan klasifikasi metode naive bayes lebih akurat dengan nilai akurasi 63,21% daripada klasifikasi KNN dengan nilai akurasi 58,94%.

Penelitian kedua dari Tempola et al., (2018) membandingkan metode KNN dan Naive Bayes pada data aktivitas status gunung berapi di Indonesia. Hasil penelitian menunjukkan bahwa rata-rata akurasi sistem menggunakan KNN adalah 63,68% dengan standar deviasi 7,47%. Sementara itu, penerapan Naive Bayes Classifier menghasilkan rata-rata akurasi sistem sebesar 79,71% dengan standar deviasi 3,55%. Kesimpulannya, akurasi sistem lebih baik ketika menggunakan Naive Bayes Classifier dibandingkan dengan KNN. Algoritma klasifikasi mempunyai kelebihan dan kelemahannya tersendiri untuk pengklasifikasian data teks (Haviluddin et al., 2022). Pemilihan algoritma KNN dalam penelitian ini didasarkan pada kemampuannya dalam menggeneralisasi dan mencapai tingkat akurasi yang cukup tinggi dalam mengklasifikasikan pola. KNN adalah metode yang digunakan untuk menganalisis data dan mengenali pola. Algoritma ini berfungsi untuk analisis klasifikasi dan regresi, serta dapat melakukan prediksi dan penilaian terhadap sistem. Tujuan dari KNN adalah memberikan nilai frekuensi kata untuk mengklasifikasikan kalimat dengan label positif dan negatif. Pada penelitian ini, peneliti melakukan crawling dataset berupa tweet atau postingan di media sosial X sebanyak 1337 postingan yang telah diklasifikasikan berdasarkan sentimen positif dan negatif. Proses crawling dataset dilakukan dengan menggunakan perangkat lunak RapidMiner untuk pengumpulan data. Penelitian ini menerapkan algoritma KNN dan Naïve Bayes sebagai model pemodelan, serta melakukan pemrograman menggunakan Python di Google Colab. Peneliti melakukan komparasi untuk menguji kedua model tersebut guna mengetahui tingkat akurasi terbaik di antara keduanya.

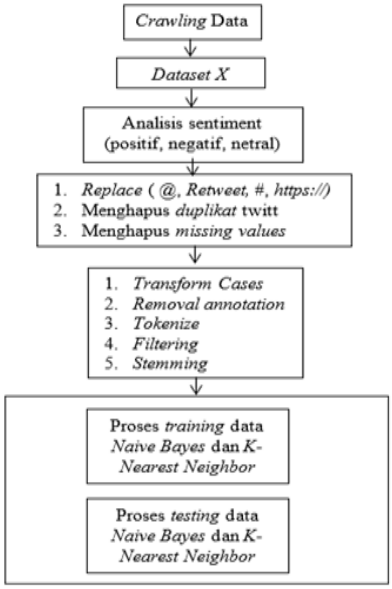
Berdasarkan hasil penelitian-penelitian sebelumnya, ada dorongan kuat untuk melakukan komparasi kedua model dengan tahapan eksplorasi preprocessing menggunakan dataset yang lebih besar pada studi kasus terkini. Penelitian ini diharapkan dapat memberikan manfaat dalam mempermudah proses klasifikasi opini di media sosial X yang berbahasa Indonesia, untuk mengetahui perbandingan tingkat akurasi klasifikasi teks menggunakan K-

Nearest Neighbor dan Naïve Bayes. Selain itu, penelitian ini juga bertujuan memberikan informasi mengenai reaksi sentimen masyarakat terhadap calon gubernur Jawa Tengah yang akan dipilih pada Pilkada tahun 2024.

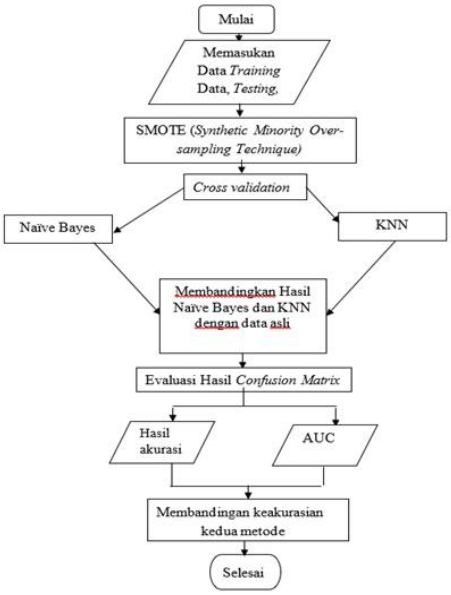
2. METODOLOGI PENELITIAN

Sumber data yang diambil dalam penelitian bersifat data sekunder karena dalam penelitian menggunakan sumber data yang berasal dari postingan di sosial media X dengan yang berkaitan dengan calon gubernur Jawa Tengah tahun 2024. Untuk mendapatkan dataset penulis memanfaatkan software RapidMiner yang telah terhubung dengan X untuk mengambil data dari bulan Januari 2024 hingga September 2024.

Proses mencari seluruh data dimulai saat menginputkan jumlah data training, jumlah data testing dan nilai k. Data training dan data testing dihitung dengan K-Nearest Neighbor dan Naive Bayes. Lalu akan dilanjutkan dengan menghitun akurasi yang diperoleh dari hasil kedua metode tersebut. Proses tersebut akan dijelaskan dalam kerangka alur pada gambar 2.



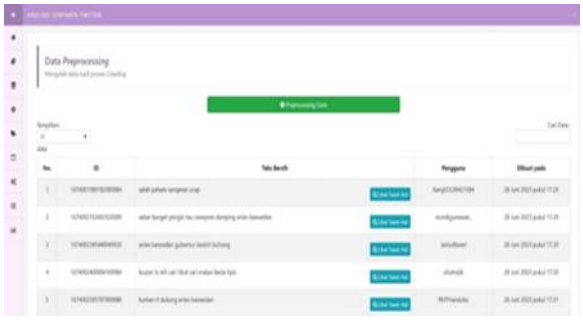
GAMBAR 1. Alur Penelitian



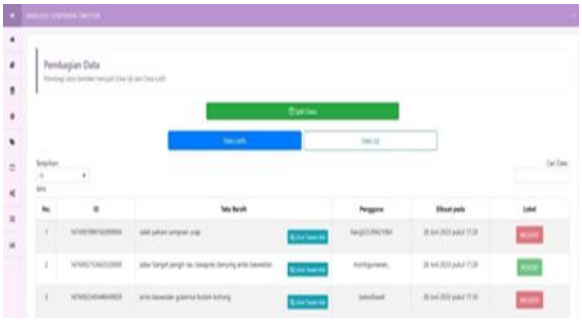
GAMBAR 2. Alur Proses Data

3. HASIL DAN PEMBAHASAN

Langkah-langkah yang diambil oleh peneliti untuk mengumpulkan data dari media sosial X melibatkan postingan teks tentang calon gubernur Jawa Tengah tahun 2024. Data yang dikumpulkan melalui scraper dari platform media sosial X akan disimpan dalam file Excel, yang kemudian akan diimpor ke dalam basis data (database) MySQL. Pemrosesan data merupakan tahap penting dalam penambangan dan analisis data, yang berfungsi untuk mengubah data mentah menjadi format yang dapat dipahami dan dievaluasi oleh komputer serta algoritma pembelajaran mesin. Data mentah ini sering kali berupa teks dengan ejaan yang tidak sesuai dengan EYD. Persiapan data sangat penting karena jika model pembelajaran mesin dilatih dengan data yang buruk atau "kotor," hasilnya biasanya akan berupa model yang kurang akurat dan kurang terlatih, sehingga tidak bermanfaat untuk analisis yang sedang dilakukan.



GAMBAR 3. Pre-Proocessing



GAMBAR 4. Pembagian Data

Pelabelan positif dan negatif adalah proses memberikan label sentimen pada data teks berdasarkan kepositifan atau kenegatifan yang dinyatakan dalam teks tersebut. Berikut cara memberi label sentimen positif dan negatif: Ketika mengategorikan sentimen positif, berarti memberikan label pada sampel teks yang menunjukkan pemikiran positif atau penuh harapan.

TABEL 1. Pelabelan

No	Teks Melalui Proses *Cleaning*	Sentimen
1	idiot habis cebong raja ngibul orang solo gue kirim sempel lo idiot	Negatif
2	tunjuk ketidaksukaannya adminnya partai layak coblos	Positif
3	Lutfi calon gubernur bodoh bohong	Negatif
4	sabar banget ini gus yasin damping lutfi	Positif
5	Pak Hendi bagus ini kapasistasnya	Positif

TABEL 2. Pengurutan Hasil Jarak

Urutan	Jarak (d(uji-I, latih-i))	Kicauan	
1	1	Uji-1	Latih-4
2	2	Uji-1	Latih-2
3	2	Uji-1	Latih-3
4	2	Uji-1	Latih-5
5	2.236	Uji-1	Latih-5
1	1	Uji-2	Latih-2
2	1.732	Uji-2	Latih-4
3	2	Uji-2	Latih-5
4	2.236	Uji-2	Latih-3
5	2.449	Uji-2	Latih-1

Pada tahap pemodelan ini, dilakukan transformasi satu set dokumen teks menjadi matriks yang menghitung frekuensi setiap token yang mewakili kata-kata (atau n-gram) dalam dokumen tersebut. Tekstual dari postingan di media sosial X diubah menjadi vektor menggunakan CountVectorizer. Tahapan ini melibatkan lima proses utama dalam pembuatan model data latih, termasuk pemilihan data latih, pembuatan kamus kata, identifikasi fitur kata, dan pembuatan vektor kosong yang diisi dengan angka yang merepresentasikan frekuensi kata.

Implementasi Performa Akurasi Sentimen Calon Gubernur Jawa Tengah di Sosial Media X Menggunakan Metode K-Nearest Neighbor

Proses klasifikasi dengan K-Nearest Neighbor (KNN), pencarian tetangga terdekat membutuhkan penentuan nilai K. Nilai K ini menunjukkan jumlah tetangga terdekat yang akan digunakan dalam proses klasifikasi. Dalam penelitian ini, beberapa nilai K yang dipertimbangkan meliputi K=3, K=5, K=7, K=9, dan K=11. Namun, nilai K yang digunakan untuk proses klasifikasi adalah K=3, yang berarti tiga tetangga terdekat akan digunakan untuk menentukan label data. Probabilitas dihitung dengan memeriksa label probabilitas yang muncul pada data K tetangga di atas. Probabilitas bahwa tes pada postingan di sosial media X berdasarkan hastag #pilkada2024, #pilkadajateng2024, #cagubjateng2024 akan diberi label positif atau negatif. Ini dapat ditentukan dengan menggunakan label_list pada model latih dan Tabel 3 data K terdekat. Tingkat probabilitas yang ditentukan seperti pada Tabel 4.

TABEL 3. Data K Terdekat

Urutan	Jarak (d(uji-I, latih-i))	Kicauan	
1	1	Uji-1	Latih-4
2	2	Uji-1	Latih-2
3	2	Uji-1	Latih-3
1	1	Uji-2	Latih-2
2	1.732	Uji-2	Latih-4
3	2	Uji-2	Latih-5

TABEL 4. Nilai Probabilitas Data Uji

Tweet (uji-i)	Tweet (latih-i)	Sentimen	Probabilitas positif	Probabilitas negatif
	Latih-4	Negatif	0	1
Uji-1	Latih-2	Negatif	0	1
	Latih-3	Positif	1	0
	Jumlah		**1**	**2**
	Latih-2	Negatif	0	1
Uji-2	Latih-4	Negatif	0	1
	Latih-5	Positif	1	0
	Jumlah		**1**	**2**

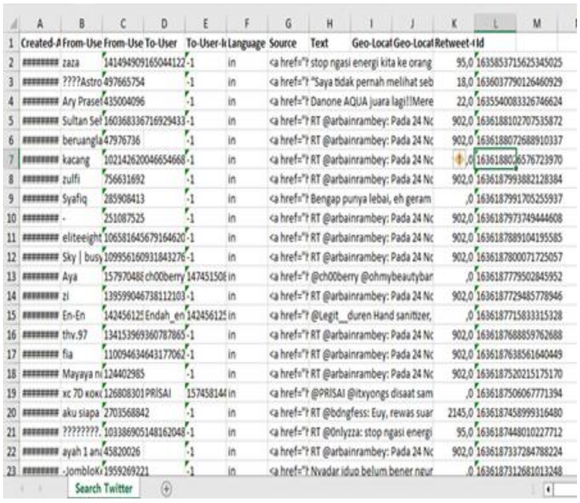
Dari hasil pengujian dan evaluasi menggunakan aplikasi yang dikembangkan dengan ekstraksi fitur CountVectorizer dan algoritma K-Nearest Neighbor (KNN) dalam tahap analisis sentimen, dapat disimpulkan

bahwa sistem berfungsi dengan baik. Hasil pengujian menunjukkan bahwa dengan menggunakan nilai K=3 pada dataset Andhika-Hendi, presisi mencapai 98,81% dan recall mencapai 99,79%.

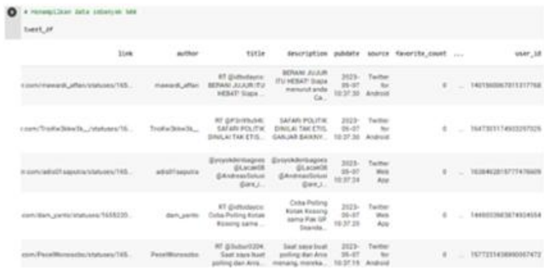
Implementasi Performa Akurasi Sentimen Calon Presiden di Sosial Media X Menggunakan Metode Naive Bayes

Crawling data yang diambil dari sosial media X dengan kata kunci yang berhubungan dengan calon gubernur Jawa Tengah tahun 2024, kemudian Andhika-Hendi dan Lutfhi-Yasin. Crawling data yang diambil dalam bentuk teks dari pengguna sosial media X. Data yang diambil pada September 2024 berjumlah 1200 teks setelah melalui tahap text preprocessing berupa opini positif dan negatif. Hasil crawling terdiri dan 12 kolom yang berisi id merupakan created_at yaitu waktu posting diunggah oleh pengguna, from_user yaitu merupakan nama pengguna X, from_user_id yang merupakan id pengguna X, to-user merupakan postingan untuk user yang lain, to-user-id merupakan nomor id user yang di-posting, language yang merupakan bahasa yang digunakan ketika membuat posting di X, source merupakan link, text merupakan isi posting di X, geo-location merupakan lokasi pengguna X, retweet-count merupakan jumlah retweet, dan id merupakan nomor id X seperti pada gambar 5.

Dataset yang digunakan dalam penelitian ini berasal dari database X dan disimpan dalam format .csv setelah dilakukan scraping menggunakan Netlytic. Data ini mencakup postingan acak yang berasal dari pemilik akun, baik itu akun asli maupun yang diduga palsu, serta media. Postingan-postingan ini berasal dari media sosial X dan terkait dengan calon gubernur Jawa Tengah tahun 2024. Total terkumpul sebanyak 1337 postingan yang masing-masing mengandung tiga kata kunci utama, yaitu #pilkada2024, #pilkadajateng2024, #cagubjateng2024.



GAMBAR 5. Hasil Crawling Data



GAMBAR 6. Data Mentah Hasil X Crawling

GAMBAR 7. Labeling Data

Data X difilter menjadi satu kolom saja, kemudian kolom yang digunakan hanya sebanyak kolom deskriptif atau review untuk memudahkan analisis proses ke tahap berikutnya. Tahapan selanjutnya adalah labelling.

Apabila nilai compound vader tersebut merupakan hasil gabungan atau hasil dari nilai rata-rata bobot sentimen. Jika nilai compound >= 0.05 maka data posting di X merupakan sentimen positif. Jika nilai compound = 0 maka data posting di X termasuk sentimen netral. Apabila nilai compound <= 0.05 maka termasuk sentimen negatif. Pada metode Naïve Bayes digunakan beberapa skenario pembagian data uji yang selanjutnya dilakukan pencarian hyperparameter alpha dengan bantuan cross validation dari tiga rasio pembagian data. Data train dan data test selanjutnya diolah menggunakan metode Naïve Bayes dan parameter terbaik yang didapatkan. Hasil akurasi dengan pembagian dataset 90%:10% dan alpha 23,4 adalah 79,29% untuk pembagian data 80%:20% dan alpha 14,4 adalah 79,80% dan untuk pembagian data 70%:30% dan alpha 29,8 menghasilkan akurasi sebesar 78,92%. Hasil akurasi terbaik adalah pada rasio dataset 80:20. Tabel 8 adalah hasil akurasi, presisi, recall dan f1-score dari beberapa rasio dataset.

Setelah mendapatkan rasio dengan nilai akurasi terbaik, kemudian dilakukan cross validation untuk mengetahui kinerja minimum dan maksimum yang didapat. Pada penelitian ini digunakan 10-folds cross validation dan hyperparameter alpha = 23,4. Tabel 9 adalah hasil dari 10-folds cross validation. 10-folds cross-validation membantu memvalidasi keandalan model dengan melihat rentang kinerja dari minimum hingga maksimum. Nilai hyperparameter alpha = 23,4 menunjukkan pengaturan optimal untuk model. Jika hasil cross-validation memiliki performa yang konsisten di semua fold (nilai akurasi tidak bervariasi secara signifikan), maka model dapat dianggap andal dan siap untuk digunakan pada data baru.

TABEL 5. Hasil Akurasi, Presisi, Recall dan f1-score

Data train : data test (%)	*alpha*	Akurasi	Presisi	*Recall*	*f1-score*
90:10	23,4	77,61%	70,24%	72,10%	71,01%
80:20	14,4	76,97%	70,47%	72,02%	71,12%
70:30	29,8	76,99%	70,06%	71,08%	70,52%

TABEL 6. Hasil 10-Folds Cross Validation

n-fold	*Accuracy*	*Presicion*	*Recall*	*f1-score*
1	78%	71%	72%	71%
2	78%	72%	74%	73%
3	79%	72%	73%	72%
4	78%	71%	73%	72%
5	76%	71%	73%	71%
6	75%	68%	69%	68%
7	77%	71%	73%	72%
8	76%	69%	69%	69%
9	78%	72%	73%	73%
10	78%	70%	72%	71%

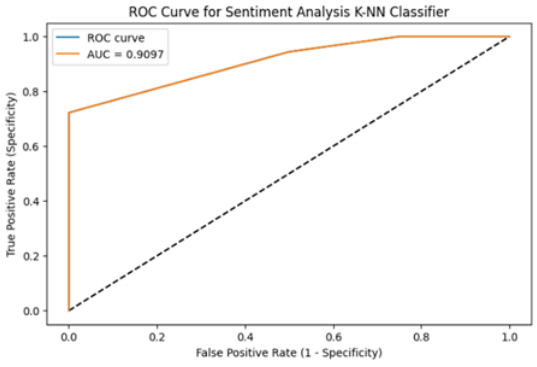
Perbandingan Antara K-Nearest Neighbor dan Naive Bayes dalam Akurasi Sentimen Analisis pada Pilihan Calon Gubernur Jawa Tengah Tahun 2024 di Sosial Media X

Berdasarkan pengujian dengan metode K-Nearest Neighbor dan Naive Bayes pada data postingan di X, didapatkan hasil kinerja dari kedua metode tersebut pada Tabel 7 berikut ini:

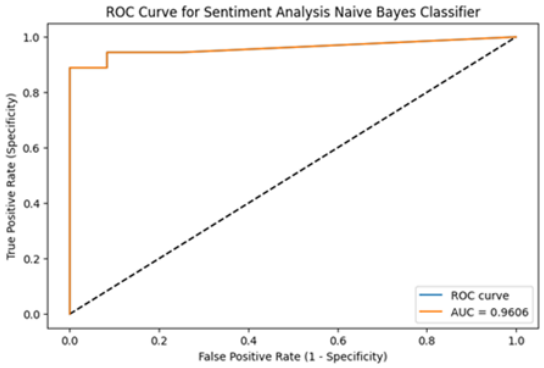
TABEL 7. Kinerja K-Nearest Neighbor dan Naive Bayes Classifier

Rasio	*K-Nearest Neighbor*				*Naive Bayes Classifier*			
	Akurasi	**Presisi**	***Recall***	***f1***	**Akurasi**	**Presisi**	***Recall***	***f1***
90:10	81,00%	78,36%	64,78%	67,21%	79,00%	70,24%	72,10%	71,01%
80:20	79,26%	75,06%	65,06%	67,09%	76,97%	70,47%	72,02%	71,12%
70:30	79,50%	77,25%	63,26%	65,17%	76,99%	70,06%	71,08%	70,52%

Dari hasil tiga pengujian yang dilakuakan peneliti dengan perbandingan rasio 90:10, 80:20, dan 70:30, dapat diketahui bahwa akurasi metode K-Nearest Neighbor lebih tinggi dibandingkan dengan Naive Bayes pada ketiga pengujian dengan data postingan di X terhadap capres 2024. Pada metode K-Nearest Neighbor, akurasi tertinggi yang didapatkan adalah 81,00%. Sementara itu, akurasi tertinggi yang didapat dengan menggunakan metode Naive Bayes adalah 79%. Hal ini dapat disebabkan karena data postingan di X sentimen terhadap capres 2024 merupakan data teks yang berbentuk kalimat dimana atributnya saling berhubungan, sehingga pada proses perhitungan algoritma K-Nearest Neighbor lebih unggul karena perhitungannya bersifat dependen sedangkan algoritma Naive Bayes perhitungannya bersifat independen.



GAMBAR 8. Kurva AUC (Area Under Curve) KNN



GAMBAR 9. Kurva AUC (Area Under Curve) Naïve Bayes

Sedangkan hasil dari algoritma klasifikasi Naïve Bayes menyatakan bahwa tingkat akurasi pada metode ini dapat dilihat pada tabel di bawah ini, yang menunjukkan bahwa data dengan kata kunci #andhika-hendi mendapatkan nilai akurasi atau sentiment positif sebesar 77%.

Pada penelitian terdahulu oleh Baharuddin, et al., (2022), klasifikasi menggunakan Naïve Bayes memberikan akurasi yang lebih baik dibandingkan dengan algoritma Maximum Entropy. Penelitian terdahulu lain dari Hananto et al., (2023), penelitian ini menggunakan dataset mengenai pendapat masyarakat Indonesia yang terdapat di X dengan hastag (#) yang mengandung Andhika-Hendi dan Lutfhi-Yasin. Tahap pra-pemrosesan data dengan melakukan seleksi komentar, pembersihan data, penguraian teks, normalisasi kalimat dan tokenisasi berdasarkan teks yang diberikan dalam bahasa Indonesia, penentuan atribut kelas, dan terakhir, mengklasifikasikan postingan Twitter dengan hashtag (#) menggunakan Naïve Bayes Classifier (NBC) dan Support Vector Machine (SVM) untuk mencapai akurasi optimasi yang optimal dan maksimal.

Penelitian terdahulu selanjutnya oleh Saputra et al., (2022), menunjukkan bahwa SVM adalah algoritma pembelajaran mesin yang berfungsi untuk menganalisis sentiment, sedangkan KNN merupakan suatu metode algoritma yang bertugas mengklasifikasikan teks dan data dengan cara mengklasifikasikan objek berdasarkan jarak terdekat dengan objek tersebut. Sedangkan penelitian terdahulu oleh Said & Manik (2022), menunjukkan bahwa model klasifikasi label tunggal Indonesia berbasis indobenchmark BERT dan RoBERTa pada fitur target dengan preprocessing menghasilkan akurasi terbaik sebesar 98.02%. Sehingga, pada penelitian terdahulu ini, mengusulkan metode deep learning dengan menggunakan model BERT (BiDirectional Encoder Representation Form Transformers) dan RoBERTa (A Robustly Optimized BERT Pretraining Approach).

Pada penelitian sekarang, dataset akan ditambahkan kombinasi algoritma machine learning lainnya atau deep learning untuk meningkatkan akurasi model. Menyediakan dataset yang lebih berkualitas dan menerapkan teknik pembersihan data yang tepat dapat membantu meningkatkan akurasi model. Memastikan data yang digunakan memiliki label yang akurat, mengatasi masalah missing values, dan menghilangkan noise dalam data dapat memberikan hasil yang lebih baik. Sehingga, selain sentiment positif dan negatif, nantinya aka nada juga faktor yang mempengaruhi dalam memilih gubernur dan wakil gubernur Jawa Tengah tahun 2024.

Pembaharuan akademik dari penelitian ini yaitu penemuan pola-pola baru, hubungan tersembunyi, dan pengetahuan yang berharga dari data yang ada. Dengan menganalisis data yang besar dan kompleks, penelitian data mining dapat membantu mengidentifikasi tren, keterkaitan, dan pola yang tidak dapat terlihat secara langsung mengenai fenomena masyarakat dalam menentukan pilihan calon kepala daerah pada Pilkada 2024 melalui sosial media X. Dataset disesuaikan dengan representasi masukan yang akan diterima oleh BERT dan RoberTa. Oleh karena itu dibutuhkan tokenizer yang bertujuan untuk menandai kalimat dan menghasilkan masukan yang sesuai. Kalimat akan diproses oleh tokenizer untuk merepresentasikan input pada BERT dan juga RobertTa. Pada rencana penelitian ini model BERT yang digunakan adalah indolem dan indobenchmark, sedangkan RoBERTa model yang digunakan adalah robert-base-indonesian. Setiap kalimat akan dipecah menjadi kata-kata menggunakan sebuah wordpiece dan akan mendapatkan ID dari kata tersebut. Setiap kata akan mendapatkan token yang sudah menjadi sistem persediaan. Setiap kalimat juga akan mendapatkan token khusus di awal dan akhir kalimat. Pada BERT token yang digunakan adalah [CLS] untuk token di awal kalimat dan [SEP] untuk token di akhir kalimat. Sedangkan di RoBERTa token yang digunakan adalah <s> untuk token di awal kalimat dan. </s> untuk token di akhir kalimat.

4. KESIMPULAN

Kesimpulan dalam penelitian ini adalah sebagai berikut:

a. Algoritma KNN memiliki performance measure dengan nilai rata-rata 81% pada dataset Andhika-Hendi pada sosial media X
b. Algoritma Naïve Bayes memiliki performance measure dengan nilai rata-rata 79% pada dataset Andhika-Hendi pada sosial media X pada n-fold 3.
c. Hasil dari pengujian mendapatkan kesimpulan algoritma KNN memiliki performance measure atau akurasi cukup tinggi dibandingkan dengan algoritma Naïve Bayes hanya sebesar 79% sementara algoritma KNN mendapat nilai rata-rata accuracy mencapai 81% yaitu dataset Andhika-Hendi, lalu precision 98,81%, recall 99,79%. Pada proporsi sentimen menunjukan sentimen positif yang diperoleh Andhika-Hendi lebih tinggi daripada calon lainnya yaitu 75%, dan Lutfhi-Yasin 15%, Sementara sentimen negatif

REFERENSI

[1] Baharuddin, T., Qodir, Z., Jubba, H., & Nurmandi, A. (2022). Prediction of Indonesian presidential candidates in 2024 using sentiment analysis and text search on Twitter. International Journal of Communication and Society, 4(2), 204–213. https://doi.org/10.31763/ijcs.v4i2.512

[2] Hananto, A. L., Nardilasari, A. P., Fauzi, A., Hananto, A., & Priyatna, B. (2023). Best Algorithm in Sentiment Analysis of Presidential Election in Indonesia on Twitter. International Journal of Intelligent Systems and Applications in Engineering, 11(6), 473–481.

[3] Haviluddin, Puspitasari, N., Burhandeny, A. E., Nurulita, A. D. A., & Trahutomo, D. (2022). Naïve Bayes and K-Nearest Neighbor Algorithms Performance Comparison in Diabetes Mellitus Early Diagnosis. International Journal of Online and Biomedical Engineering, 18(15), 202–215. https://doi.org/10.3991/ijoe.v18i15.34143

[4] Hidra Amnur, A. K. Vadreas, and M. Ridwan, "Aplikasi Pendeteksi Kematangan Tanaman Menggunakan Metode Transformasi Ruang Warna HSI (Hue, Saturation, Intensity) dan K-NN (K- Nearest Neighbor)", jitsi, vol. 5, no. 4, pp. 161 -167, Dec. 2024.

[5] Hozairi;, Anwari;, & Alim, S. (2021). Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive BayeS. Network Engineering Research Operation, 6(2), 133–144. https://doi.org/10.21107/NERO.V6I2.237

[6] Ilić, M., Srdjević, Z., & Srdjević, B. (2022). Water quality prediction based on Naïve Bayes algorithm. Water Science and Technology, 85(4), 1027–1039. https://doi.org/10.2166/wst.2022.006

[7] Novianti, N., Zarlis, M., & Sihombing, P. (2022, April). Penerapan Algoritma Adaboost Untuk Peningkatan Kinerja Klasifikasi Data Mining Pada Imbalance Dataset Diabetes | Novianti | JURNAL MEDIA INFORMATIKA BUDIDARMA.

[8] Nuqoba, B., & Djunaidy, A. (2014). Algoritma Prediksi Outlier Menggunakan Border Solving Set. Jurnal Informatika Mulawarman, 9(3), 10.

[9] Rosso, G. A. (2019). Milton. William Blake in Context, (September), 184–191. https://doi.org/10.1017/9781316534946.021

[10] Said, F., & Manik, L. P. (2022). Aspect-Based Sentiment Analysis on Indonesian Presidential Election Using Deep Learning. Paradigma - Jurnal Komputer Dan Informatika, 24(2), 160–167. https://doi.org/10.31294/paradigma.v24i2.1415

[11] Saputra, N., Nurbagja, K., & Turiyan, T. (2022). Sentiment Analysis of Presidential Candidates Anies Baswedan and Ganjar Pranowo Using Naïve Bayes Method. Jurnal Sisfotek Global, 12(2), 114. https://doi.org/10.38101/sisfotek.v12i2.552

[12] Syarifuddinn, M. (2020). Analisis Sentimen Opini Publik Mengenai Covid-19 Pada Twitter Menggunakan Metode Naïve Bayes Dan KNN. INTI Nusa Mandiri, 15(1), 23–28. https://doi.org/10.33480/INTI.V15I1.1347

[13] Tempola, F., Muhammad, M., & Khairan, A. (2018, October). Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation | Tempola | Jurnal Teknologi Informasi dan Ilmu Komputer.