Jurnal KomtekInfo

https://jkomtekinfo.org/ojs

2024 Vol. 11 No. 4 Hal: 323-331 e-ISSN: 2502-8758

Metode Support Vector Machine dan Naïve Bayes untuk Analisis Santimen Ibu Kota Nusantara

Muhamad Rafi Akbar✉, Sarjon Defit, Sumijan

Fakultas Ilmu Komputer, Universitas Putra Indonesia YPTK Padang, 25221, Indonesia

muhamadrafiakbar21@gmail.com

Abstract

The relocation of the Indonesian Capital City (IKN) has raised various pros and cons. Starting from the selection of the location, the ratification of the Law which is considered too hasty, and recently the government has also invited the Indonesian people to crowdfund to build the Indonesian Capital City. This study aims to compare the effectiveness of two approaches to classification: Support Vector Machine and Naïve Bayes, in analyzing opinion sentiment towards the Indonesian Capital City based on social media data such as TikTok. Opinion sentiment analysis is very important to understand public views on various aspects of the Indonesian Capital City. The data used will involve opinions that develop on social media regarding the Indonesian Capital City. The methods used in this study are Support Vector Machine and Naïve Bayes. The research methodology includes data collection, preprocessing, data sharing, training Naïve Bayes and SVM models, evaluation, and statistical analysis to compare the performance of the two models. The dataset consists of 1529 comments taken from the TikTok application. The final results of the evaluation carried out can be seen in the comparison between the Support Vector Machine and Naïve Bayes methods based on the level of accuracy obtained by each method. Support Vector Machine obtained an aura rate of 98%, where the accuracy rate is lower than the accuracy rate of the Naive Bayes method with a percentage of 92%. Based on the findings of the analysis, the procedure using the Support Vector Machine method showed better results than the Naïve Bayes method in measuring sentiment toward the Indonesian Capital.

Keywords: *Nusantara Capital City, TikTok, Sentiment Analysis, Support Vector Machine, Naïve Bayes*

Abstrak

Pemindahan Ibu Kota Nusantara (IKN) menimbulkan berbagai macam pro dan kontra. Mulai dari pemilihan lokasi, pengesahan Undang – Undang yang dinilai terlalu terburu-buru, dan akhir-akhir ini pemerintah juga mengajak masyarakat Indonesia untuk melakukan crowd funding untuk membangun Ibu Kota Nusantara. Penelitian ini bertujuan untuk membandingkan efektivitas dua pendekatan untuk klasifikasi: Support Vector Machine dan Naïve Bayes, dalam menganalisis sentimen opini terhadap Ibu Kota Nusantara berdasarkan data media sosial seperti tiktok. Analisis sentimen opini sangat penting untuk memahami pandangan publik mengenai berbagai aspek Ibu Kota Nusantara. Data tersebut yang digunakan akan melibatkan opini yang berkembang di media sosial mengenai Ibu Kota Nusantara. Metode yang digunakan dalam penelitian ini adalah Support Vector Machine dan Niave Bayes. Metodologi penelitian mencakup pengumpulan data, preprocessing, membagi data, pelatihan model Naïve Bayes dan SVM, evaluasi, serta analisis statistik untuk membandingkan kinerja kedua model. Dataset terdiri dari 1529 komentar yang diambil dari aplikasi Tiktok. Hasil akhir dari evaluasi yang dilakukan dapat dilihat perbandingan anatara metode Support Vector Machine dengan Naive Bayes berdasarkan tingkat akurasi yang diperolch oleh masing-masing metode. Support Vector Machine memperoleh tingkat aurasi 98%, di mana tingkat akurasinya lebih rendah daripada tingkat akurasi metode Naive Bayes dengan persentase 92%. Berdasarkan temuan analisis, prosedur yang menggunakan metode Support Vector Machine menunjukkan hasil yang lebih baik dibandingkan metode Naïve Bayes dalam mengukur sentimen terhadap Ibu Kota nusantara.

Kata kunci: Ibu Kota Nusantara, *Tiktok*, Analisis Sentimen, *Support Vector Machine*, *Naïve Bayes*

1. Pendahuluan

Knowledge Discovery in Database (KDD) atau yang dikenal dengan Data Mining adalah proses pengumpulan data yang bertujuan untuk menemukan pola, pengetahuan, dan juga informasi. Namun, pasti ada algoitma atau teknik yang digunakan untuk menemukan pola tersebut. Output yang dapat dihasilkan dari proses data mining digunakan sebagai pilihan untuk pengambilan keputusan [1]. Tujuan dari Knowledge Discovery in Database (KDD) dan data mining adalah untuk menerapkan metode saintifik pada data mining. Data Mining (DM) adalah inti dari proses KDD, melibatkan kesimpulan dari algoritma yang mengeksplorasi data, mengembangkan model dan menemukan pola yang sebelumnya tidak diketahui. Tujuan KDD dan data mining adalah untuk menggali informasi tersembunyi dari sebuah basis data yang sangat besar [2]. Teks mining adalah bagian khusus dari pengolahan data, yang berarti menambang data berupa teks. Proses ini biasanya dilakukan dengan menggunakan dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dokumen sehingga

Diterima: 25-08-2024 | Revisi: 27-09-2024 | Diterbitkan: 31-12-2024 | doi: 10.35134/komtekinfo.v11i4.579

dapat dilakukan analisis hubungan antar dokumen. Pengolahan teks dapat dilakukan dengan komputer, dan menghasilkan temuan informasi baru [3].

Text Mining adalah proses inovasi akan informasi atau isu terkini yang sebelumnya tidak terungkap menggunakan mekanisme dan menganalisa data pada jumlah besar. Proses dalam menganalisa keseluruhan atau sebagian unstructured text, text mining membuat asosiasikan satu bagian text dengan lainnya berdasarkan aturan tertentu. Hasil yang diharapkan adalah kata baru yang tidak terungkap jelas sebelumnya [4].

Klasifikasi adalah teknik untuk mendapatkan fungsi untuk membedakan jenis kategori atau kelas data. Banyak teknik yang tersedia dalam bidang data mining yang dapat digunakan untuk mengolah sejumlah besar data menjadi informasi bermanfaat [5]. Tujuan klasifikasi adalah untuk memprediksi atau memperkirakan kelas dari data baru yang belum memiliki merek. Untuk mencapai tujuan ini, perlu dibuat cara untuk membedakan kelas data dengan metode tertentu [6]. Salah satunya yaitu menggunakan Support Vector Machine (SVM) yang merupakan salah satu algoritma pengajaran mesin yang paling banyak digunakan untuk klasifikasi.

Periode dalan waktu sepuluh tahun terakhir, SVM telah berkembang menjadi alat yang kuat untuk pola klasifikasi dengan tingkat keberhasilan yang tinggi saat digunakan di berbagai industri. Karena SVM dapat menangani berbagai masalah pembelajaran, banyak komunitas pembelajaran mesin tertarik untuk mempelajari dan mengembangkannya. Metode pembelajaran mesin yang dikenal sebagai SVM bertujuan untuk menemukan hyperplane terbaik yang dapat memisahkan dua kelas pada ruang input. Dengan menggunakan data training, algoritma klasifikasi SVM membuat model klasifikasi. Model ini digunakan untuk memprediksi kelas data baru yang belum pernah ada sebelumnya, yang disebut data testing [7].

Penggunaan klasifikasi Naive Bayes adalah metode klasifikasi statistik yang digunakan untuk memprediksi kemungkinan keanggotaan kelas tertentu, menghitung kemungkinan untuk suatu hipotesis, dan menghitung kemungkinan kelas dari setiap kelompok atribut yang ada, serta menentukan kelas mana yang paling optimal [8]. Penelitian Joshua Muliawan dan Erick Dazki di dapat hasil penelitian analisis sentimen pemindahan ibu kota negara indonesia menggunakan tiga algoritma: naïve bayes, knn, dan random forest didapatkan nilai akurasi metode Algoritma Naïve Bayes Classifier yang didapati nilai keakuratan sebesar 65.26%, Algoritma K-Nearest Neighbor sebesar 58.25%, serta Algoritma Random Forest sebesar 45.05% [9].

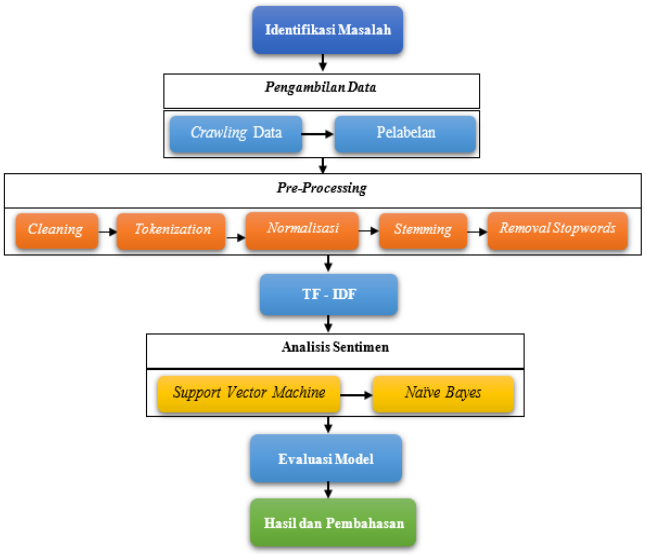
Sedangkan penelitian lain yang dilakukan oleh Sumayah dkk, membahas tentang Analisis Sentimen Masyarakat Indonesia Terhadap Metaverse menggunakan Algoritma Support Vector Machine dengan jumlah data sebanyak 2504 data mendapatkan hasil akurasi tertinggi sebesar 81% [10].

Penelitian lain yang dilakukan oleh Pramana dkk, mengenai analisis sentimen terhadap pemindahan ibu kota negara Indonesia dengan membandingkan dua algoritma, yaitu Naïve Bayes, dan K-Nearest Neightbors. Dari pengujian ini, algoritma Naïve Bayes memiliki skor akurasi 63.09% dan algoritma K-Nearest Neightbors mendapatkan skor akurasi 6 sebesar 69.23% [11]. Penelitian selanjutnya dilakukan oleh Subarkah dkk, membahas tentang Sentiment Analysis On Reviews Of Women's Tops On Shopee Marketplace Using Naive Bayes Algorithm dapat disimpulkan hasil pengujian menggunakan naïve bayes menunjukkan nilai akurasi sebesar 89% [12]. Penelitian selanjutnya dilakukan oleh Putri Elisa, dan Auliya Rahman Isnain, membahas tentang Comparison Of Random Forest, Support Vector Machine And Naïve Bayes Algorithms To Analyze Sentiment Towards Mental dapat disimpulkan hasil pengujian model SVM menunjukkan akurasi sebesar 86.11%, model Random Forest menunjukkan akurasi sebesar 82.55%, sedangkan model Naive bayes menunjukkan akurasi sebesar 78.19%. Oleh karena itu, dapat disimpulkan bahwa SVM memiliki performa yang paling baik dalam mengklasifikasikan tweet yang mengandung stigma kesehatan mental [13]. Penelitian lainnya oleh Sarimole dan Septian juga meneliti tentang Analisis Sentimen Masyarakat Terhadap Isu Penundaan Pemilu 2024 Pada Twitter Dengan Metode Naive Bayes Dan Support Vector Machine. Didapati hasil pengujian model SVM menunjukkan akurasi sebesar 91.61%, sedangkan model Naive Bayes menunjukkan akurasi sebesar 98.80%, yang berarti model Naïve Bayes lebih akurat dibanding model Support Vector Machine [14].

Melalui pelaksanaan penelitian ini, diharapkan dapat mengungkap pemahaman tentang respons masyarakat terhadap pemindahan Ibu Kota Negara Indonesia, apakah cenderung bersifat positif, negatif, atau netral. Selain itu, tujuan penelitian ini adalah untuk menemukan kata-kata yang sering muncul dalam setiap sentimen, baik itu negatif, negatif, atau netral. Dengan perbandingan ini, diharapkan dapat digunakan untuk membedakan kedua metode algoritma Support Vector Machine dan Naive Bayes Classifier, yang lebih akurat dalam menilai sentimen topik pembicaraan, terutama yang berkaitan dengan pemindahan atau relokasi ibu kota negara ke IKN.

2. Metodologi Penelitian

Metodologi penelitian bertujuan untuk membantu penulis menyelesaikan masalah penelitian dan mencapai tujuan. Agar penelitian berhasil, desain penelitian diperlukan. Desain penelitian berikut menunjukkan proses pengklasifikasian dokumen teks dengan menggunakan metode Support Vector Machine dan Niave Bayes disajikan pada Gambar 1.



Gambar 1. Kerangka Penelitian

Diagram alur ini menggambarkan proses analisis sentimen pada teks secara umum. Dimulai dari identifikasi masalah yang ingin dipecahkan, kemudian dilanjutkan dengan pengumpulan data teks yang relevan. Data tersebut kemudian diolah melalui berbagai tahap pra-pemrosesan untuk membersihkan dan mempersiapkan data agar siap untuk dianalisis sentimen. Terakhir, data yang sudah bersih akan dianalisa menggunakan algoritma tertentu seperti Support Vector Machine atau Naive Bayes untuk menentukan sentimen (positif, negatif, atau netral) dari teks tersebut. Hasil analisis kemudian dievaluasi dan dibahas untuk menarik kesimpulan.

1. Tahap Identifikasi Masalah

Kualitas penelitian dipengaruhi oleh masalah penelitian, sebagai upaya untuk menjelaskan masalah dan dapat mengukur penjelasan masalah tersebut identifikasi masalah dilakukan. Proses identifikasi masalah merupakan pendefinisian dari masalah yang ada dalam sebuah penelitian. Dengan kata lain, tahap ini layaknya pondasi yang kokoh bagi seluruh bangunan penelitian

2. Tahap Pengambilan Data

Proses crawling data komentar dari media sosial Tiktok dari unggahan pengguna Tiktok dengan menggunakan Scraper dengan kata kunci Ibu Kota Nusantara (IKN). Data yang terkumpul kemudian akan diolah dan dianalisis untuk menghasilkan temuan-temuan penelitian yang valid dan reliabel. Bahasa pemrograman yang di gunakan dalam penelitian ini menggunakan bahasa pemrograman Python yang selanjutnya disimpan dalam format .csv, dari data yang didapat akan dilanjutkan dengan tahap preprocessing.

3. Tahap Pre-Processing

Pada tahap Preprocessing berfungsi untuk mengatasi kesalahan dalam mengambil ciri atau atribut yang dapat mengurangi kinerja analisis sentimen. Data yang set yang telah dikumpulkan sebelumnya mash memiliki struktur data yang tidak terstruktur dan tidak beraturan, maka preprocessing diperlukan sebelum dataset diuji dengan sebuah model. Preprocessing merupakan tahap yang dilakukan guna membersikan data dari noise dan mengubah data menjadi data yang terstruktur [15]. Tahapan dalam preprocessing dalam penelitian ini adalah :

a. Case folding adalah proses menyamakan huruf pada teks menjadi huruf kecil dimana tidak setiap teks akan konsisten dalam menggunakan huruf kapital.
b. Cleansing merupakan proses menganalisa kualitas data dengan cara mengubah, memodifikasi, atau menghapus data-data yang dianggap tidak lengkap.
c. Normalisasi adalah proses menghilangkan tanda baca, angka, simbol, link URL dan username di dalam teks.
d. Stopword Removal adalah langkah dimana frasa yang tidak penting dalam penambangan teks untuk divisi apa pun dihilangkan.
e. Stemming adalah tahapan yang digunakan dalam pemotongan awal atau akhir kata dengan memperhatikan awalan umum dan sufiks, yang dapat ditemukan dalam kata.
f. Langkah selanjutnya yaitu tokenisasi, dimana pada tahapan ini akan menghilangkan tanda baca yang tidak diperlukan dan memotong teks menjadi kata, simbol, karakter atau tanda baca, shingga menjadi token yang dapat dianalisis

4. Tahap TF-IDF

TF-IDF adalah proses pembobotan pada masing-masing kata. Pembobotan TF-IDF dinilai penting, hal ini dikarenakan apabila suatu kata lebih sering muncul dalam suatu dokumen maka nilai kontribusinya akan semakin besar, akan tetapi jika hal tersebut terjadi pada beberapa dokumen makan kontribusi yang dimiliki akan lebih kecil. TF-IDF menggunakan rumus menghitung nilai bobot dokumen :

$$W_{at} = TF_{at} \times IDF_{ft} \quad (1)$$

Keterangan :

W_{at} = Nilai dokumen ke-d pada kata ke-t
Tf_{at} = Jumlah kata yang dicari dalam suatu dokumen
IDF_{ft} = Inverse dokument frequency $\left(\log\left(\frac{D}{df}\right)\right)$
D = jumlah dokumen
df = Jumlah dokumen yang mengandung kata

TF (Term Frequency): Menunjukkan seberapa sering suatu kata muncul dalam sebuah dokumen. Semakin sering suatu kata muncul, semakin tinggi nilai TF-nya, dan semakin penting dianggap kata tersebut dalam dokumen itu.

IDF (Inverse Document Frequency): Menunjukkan seberapa unik suatu kata dalam keseluruhan kumpulan dokumen. Semakin jarang suatu kata muncul dalam keseluruhan dokumen, semakin tinggi nilai IDF-nya, dan semakin penting dianggap kata tersebut dalam membedakan dokumen satu dengan yang lain.

TF × IDF: Dengan mengalikan TF dan IDF, kita mendapatkan skor TF-IDF yang mengindikasikan pentingnya suatu kata dalam konteks dokumen dan seluruh kumpulan dokumen. Kata-kata yang sering muncul dalam dokumen tertentu tetapi jarang muncul di dokumen lain akan memiliki skor TF-IDF yang tinggi, sehingga dianggap sebagai kata kunci yang penting untuk mengkarakterisasi dokumen tersebut.

5. Tahap Analisa Sentimen

Pada tahapan ini penggunaan metode untuk melakukan analisis sentiment dilakukan. Metode yang digunakan adalah Support Vector Machine dan Naïve Bayes. Penggunaan dua metode sendiri selain untuk melihat opini public terhadap Ibu Kota Nusantara adalah untuk menguji dan membandingkan penggunaan kedua metode tersebut.

6. Tahap Evalusi Model

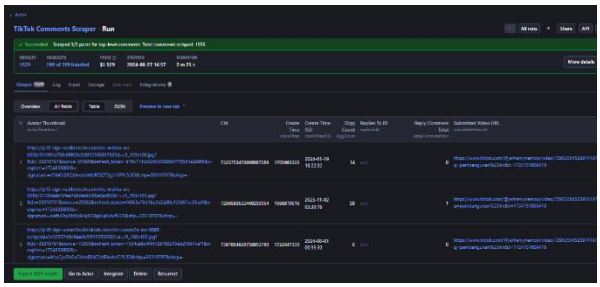
Setelah proses klasifikasi yang dilakukan di atas maka tahap selanjutnya yaitu tahap evaluasi klasifikasi. Tahap ini akan dilakukan pengujian menggunakan confusion matrix dengan matrik ukuran 3x3.

7. Hasil dan Pembahasan

Setelah setiap tahap dan proses dilakukan, selanjutnya adalah tahap visualisasi. Pada penelitian ini, untuk tahap visualisasi dilakukan dengan menggunakan library Matplotlib dan Wordcloud. Output dari visualisasi ini adalah berupa gambar histogram yang juga akan menampilkan hasil akurasi persentase dari polaritas setiap sentimen yang dihasilkan. Sedangkan untuk visualisasi wordcloud menampilkan kata yang sering muncul pada setiap sentimennya.

3. Hasil dan Pembahasan

Pada penelitian analisis sentimen ini menggunakan data komentar Tiktok *crawling* menggunakan *scraping* web bernama Apify.com dengan jumlah 1529 data yang sudah melalui proses pembersihan data. Dengan menggunakan metode Naive Bayes Classifier, yang akan dibandingkan dengan metode Support Vector Machine (SVM), akan dibahas pemindahan Ibu Kota Negara, Nusantara. Tujuannya untuk mengetahui hasil perbandingan keberhasilan nilai accuracy, precision, dan recall. dengan beberapa langkah preprocessing, validasi dan evaluasi proses preprocessing, dan label sentimen untuk data tweet, yang dibagi menjadi label sentimen positif dan negatif. Gambar 2 menunjukkan proses pengambilan data komentar Tiktok.



Gambar 2 Crawling Data menggunakan Scraper APIFY

Gambar 2 tersebut menunjukkan tampilan antarmuka (interface) dari sebuah alat atau perangkat lunak yang digunakan untuk mengambil data komentar dari sebuah situs APIFY. Website ini telah berhasil menjalankan tugasnya dengan mengambil data dari sejumlah URL tertentu. Hasil pengambilan data tersebut ditampilkan dalam bentuk tabel yang berisi informasi seperti URL video, jumlah komentar, waktu pembuatan, dan detail lainnya. Website ini tampaknya memiliki fitur untuk menyimpan dan mengelola data yang telah diambil. Intinya, gambar ini menggambarkan proses pengambilan data komentar secara otomatis dari sebuah platform online.

1. Pelabelan Data

Data harus dilabelkan dan dibagi menjadi 3 kelas sentimen yaitu kelas sentimen positif, negatif dan kelas sentimen netral berdasarkan kata – kata yang terdapat pada data komentar tiktok, dimana polarity 1 itu berlabel positif, polarity -1 berlabel negatif, dan polarity 0 berlabel netral. Contoh data yang sudah dilabelkan dapat dilihat pada Gambar 3.

	createTimeISO	uniqueId	text	label
0	2024-01-19T11:22:02.000Z	panduhabibi	jokowi presiden is the best..	Positive
1	2023-11-01T20:30:16.000Z	lemme.go1	80% investor	Neutral
2	2024-07-31T17:35:30.000Z	heiligo_007	New Atlantis	Neutral
3	2024-07-16T03:35:16.000Z	user94461265294540	lost in junglr	Negative
4	2024-08-16T15:26:17.000Z	venyoliv	@Nano_orin782 @@Phian @Gerry bel	Neutral
...	...	...	...	...
1524	2024-03-04T06:55:42.000Z	hermynsuprajuningsih	keren banget 👍👍😍😍	Positive
1525	2024-03-03T04:01:44.000Z	arrayah88	ini kan gambarnya yg kau tunjukkan . asli nya ...	Positive
1526	2024-03-03T03:56:12.000Z	rifnaz...rrq	🤣🤣🤣🤣	Neutral
1527	2024-03-03T03:54:14.000Z	blackadam878	impian anak cucu kita utk tinggal disana dgn b...	Negative
1528	2024-03-03T03:50:18.000Z	blackadam878	semoga keadaannya spt ini spy jd ibukota negar...	Positive

1529 rows × 4 columns

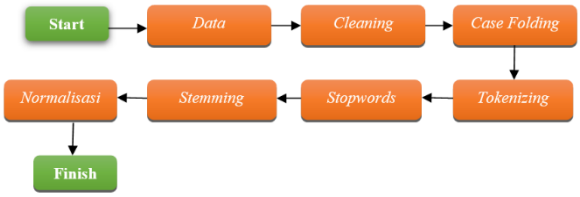
Gambar 3 Pelabelan Data

Gambar 3 tersebut menampilkan sebagian data yang diperoleh dari proses pengambilan komentar secara otomatis dari sebuah platform online. Data yang diambil

mencakup informasi penting seperti waktu komentar dibuat, ID unik pengguna yang membuat komentar, isi komentar itu sendiri, serta label sentimen yang diberikan pada komentar tersebut (positif, negatif, atau netral). Data ini disajikan dalam bentuk tabel yang mudah dibaca, dengan setiap baris mewakili satu komentar dan setiap kolom mewakili satu jenis informasi. Tabel ini memberikan gambaran yang jelas mengenai jenis data yang berhasil diambil dan dapat digunakan untuk analisis lebih lanjut. Dengan adanya data ini, kita dapat melakukan berbagai analisis, seperti menganalisis sentimen umum terhadap suatu topik, mengidentifikasi pengguna yang paling aktif, atau bahkan melakukan pemodelan prediksi sentimen pada komentar baru.

2. Pre-Processing Data

Tahapan preprocessing data merupakan bagian krusial pada proses analisis sentimen. Proses ini dilakukan agar nanti data dapat terstruktur yang kemudian dapat digunakan dalam menganalisis sentimen. Sebagaimana yang telah digambarkan pada metodologi penelitian, proses dari preprocessing data meliputi kegiatan Cleanning, Case Folding, Normalisasi, Stopword Removal, Stemming dan Tokenisasi tersaji pada Gambar 4.



Gambar 4 Langkah Preprocessing

Dagram alur ini menggambarkan langkah-langkah umum dalam proses pra-pemrosesan teks. Proses ini merupakan tahap awal yang sangat penting sebelum data teks dapat digunakan untuk analisis lebih lanjut, seperti analisis sentimen atau pengelompokan teks. Secara garis besar, proses ini dimulai dengan pengumpulan data teks mentah, kemudian dilakukan pembersihan data (cleaning), mengubah semua huruf menjadi huruf kecil (case folding), normalisasi, penghilangan kata-kata stop (stopwords), dan terakhir adalah tokenisasi.

	createTimeISO	uniqueId	text	label	clean_teks
0	2024-01-19T11:22:02.000Z	panduhabibi	jokowi presiden is the best..	Positive	[jokowi, presiden, is, best]
1	2023-11-01T20:30:16.000Z	lemme.go1	80% investor	Neutral	[investor]
2	2024-07-31T17:35:30.000Z	heiligo_007	New Atlantis	Neutral	[new, atlantis]
3	2024-07-16T03:35:16.000Z	user94461265294540	lost in junglr	Negative	[lost, in, junglr]
4	2024-08-16T15:26:17.000Z	venyoliv	@Nano_orin782 @@Phian @Gerry bel	Neutral	[]
...	...	...	...	...	...
1524	2024-03-04T06:55:42.000Z	hermynsuprajuningsih	keren banget 👍👍😍😍	Positive	[keren, banget]
1525	2024-03-03T04:01:44.000Z	arrayah88	ini kan gambarnya yg kau tunjukkan . asli nya ...	Positive	[gambar, kau, asli, debu, hutan]
1526	2024-03-03T03:56:12.000Z	rifnaz...rrq	😂😂😂😂	Neutral	[]
1527	2024-03-03T03:54:14.000Z	blackadam878	impian anak cucu kita utk tinggal disana dgn b...	Negative	[impi, anak, cucu, tinggal, aneka, ragam, suku...
1528	2024-03-03T03:50:18.000Z	blackadam878	semoga keadaannya spt ini spy jd ibukota negar...	Positive	[ibukota, negara, eksotik, dunia, rumah, lingk...

1514 rows × 5 columns

Gambar 5 Hasil Proses Preprocessing

Gambar 5 menampilkan sebagian data hasil pengambilan komentar dari sebuah platform online. Tiap baris pada tabel merepresentasikan satu komentar yang berisi informasi seperti waktu komentar dibuat, ID unik pengguna, isi komentar, label sentimen (positif, negatif, atau netral), dan hasil pembersihan teks (clean_teks). Kolom "clean_teks" menunjukkan hasil dari proses pembersihan data teks yang bertujuan untuk menghilangkan karakter-karakter yang tidak perlu, seperti tanda baca dan emoji, serta mengubah semua kata menjadi huruf kecil. Proses ini dilakukan untuk mempersiapkan data teks agar siap untuk diproses lebih lanjut, misalnya untuk analisis sentimen atau pembuatan model bahasa. Dengan adanya data yang telah dibersihkan ini, kita dapat melakukan berbagai analisis teks seperti menghitung frekuensi kata, melakukan pengelompokan topik, atau membangun model prediksi sentimen.

3. Pembobotan TF- IDF

Langkah pembobotan TF-IDF yaitu mekanisme yang mengganti data teks. Proses tersebut akan membentuk data numerik untuk menghitung jenis setiap kata atau fitur. Simulasi Pembobotan TF-IDF terdapat pada Gambar 6.

```
[ ] from sklearn.feature_extraction.text import TfidfVectorizer
    tf_idf = TfidfVectorizer(ngram_range=(1,2))
    tf_idf.fit(X)
```

```
TfidfVectorizer
TfidfVectorizer(ngram_range=(1, 2))
```

```
[ ] #Jumlah Fitur
    print(len(tf_idf.get_feature_names_out()))
```

```
8214
```

Gambar 6 Pembobotan TF-IDF

Gambar 6, Kode di atas dimaksudkan untuk mengubah teks menjadi representasi TF-IDF dan melakukan transformasi pada data latih dan data uji yang akan diuji.

ertama, kode mengimpor kelas TfidfVectorizer dari pustaka scikit-learn yang digunakan untuk menghitung skor TF-IDF. Kemudian, sebuah objek TfidfVectorizer dibuat dengan parameter ngram_range=(1,2) yang artinya akan dipertimbangkan kata tunggal (unigram) dan kombinasi dua kata (bigram). Terakhir, metode fit diterapkan pada data teks X untuk menghitung skor TF-IDF. Hasil dari proses ini adalah sebuah matriks numerik yang merepresentasikan pentingnya setiap kata atau kombinasi kata dalam setiap dokumen. Kode selanjutnya menghitung jumlah fitur (kata atau kombinasi kata) yang dihasilkan dari proses transformasi tersebut. Hasilnya, terdapat 8214 fitur yang ditemukan dalam data teks.

4. Pemisahan Data Latih dan Data Uji

Setelah diperoleh data, maka data tersebut untuk dipecah menjadi 2, data uji dan data latihan. Pada langkah analisis ini, banyak data uji 20% dan banyak data latih 80%. Pemisahan data latih dan data uji merupakan langkah penting dalam membangun model machine learning. Data latih digunakan untuk melatih model agar dapat belajar mengenali pola-pola dalam data. Setelah model dilatih, data uji digunakan untuk mengukur kinerja model pada data yang belum pernah dilihat sebelumnya. Dengan membagi data, kita dapat menghindari overfitting, yaitu kondisi di mana model terlalu cocok dengan data latih sehingga performanya buruk pada data baru. Proses ini memungkinkan kita untuk mengevaluasi seberapa baik model kita akan bekerja di dunia nyata.

5. Klasifikasi Support Vector Machine

Pada klasifikasi SVM digunakan algoritma kernel tunggal agar mengetahui poin ketelitian yang di hasilkan. Sesudah melaksanakan pemeriksaan ketelitian, pada penjumlahan ketelitian yang dilaksanakan, dihasilkan. hasil klasifikasi pola mampu meramalkan dengan ketelitian yaitu 97.52%. keputusan bisa diketahui pada Gambar 7.

```
Jumlah prediksi benar   : 1181
Jumlah prediksi salah   : 30
Akurasi pengujian       : 97.52270850536746 %
```

Gambar 7 Akurat SVM

Gambar 7 tersebut menunjukkan hasil evaluasi kinerja suatu model machine learning. Model ini telah melakukan prediksi terhadap suatu data, dan hasilnya menunjukkan bahwa dari total data yang diprediksi, sebanyak 1181 prediksi benar dan 30 prediksi salah. Berdasarkan hasil tersebut, dapat disimpulkan bahwa akurasi model ini adalah sebesar 97,52%. Artinya, model ini mampu memprediksi dengan benar hampir 98% dari data yang diberikan.

6. Klasifikasi Naïve Bayes

Pada klasifikasi NB digunakan kegunaan MultinomialNB. MultinomialNB dipakai untuk mengklasifikasi NB dan bisa mengatur data dalam bentuk teks. Sesudah melaksanakan uji presisi, berlandaskan penjumlahan ketelitian yang dilaksanakan, dihasilkan bahwa pola akan meramalkan dengan ketelitian yaitu 91.65%. keputusan bisa diketahui pada gambar 8.

```
Jumlah prediksi benar   : 1110
Jumlah prediksi salah   : 101
Akurasi pengujian       : 91.6597853014038 %
```

Gambar 8 Akurat Naive Bayes

Gambar 8 di atas menunjukkan hasil evaluasi kinerja dari sebuah model machine learning. Dari total data yang diuji, model ini berhasil memprediksi dengan benar sebanyak 1110 data dan salah sebanyak 101 data. Dengan kata lain, model ini memiliki akurasi sebesar 91,66%. Artinya, model ini mampu memprediksi dengan tepat sekitar 91,66% dari keseluruhan data yang diberikan.

7. Evaluasi Confusion Matrics

Confusion matrics yaitu alat pengukuran dalam bentuk matrics yang dipakai untuk memperoleh tingkat keakuratan klasifikasi kelas-kelas berdasarkan algoritma yang digunakan. Dengan menggunakan confusion matrix, kita dapat menghitung berbagai metrik seperti akurasi, precision, recall, dan F1-score untuk menilai seberapa baik model dalam mengklasifikasikan data. Kesimpulan perbandingan dapat di hasilkan dari kedua metode dapat diketahui pada Gambar 9.

Classification report:

	precision	recall	f1-score	support
Negatif	1.00	0.97	0.98	490
Netral	0.94	0.99	0.97	438
Positif	0.99	0.95	0.97	283
accuracy			0.98	1211
macro avg	0.98	0.97	0.97	1211
weighted avg	0.98	0.98	0.98	1211

Gambar 9 Hasil Evaluasi SVM

Gambar 9 menyajikan evaluasi kinerja model dalam mengklasifikasikan data menjadi tiga kategori: negatif, netral, dan positif. Nilai-nilai dalam gambar menunjukkan metrik-metrik seperti presisi, recall, F1-score, dan akurasi untuk setiap kelas. Misalnya, untuk kelas negatif, model memiliki presisi 1.00 yang berarti semua prediksi positif untuk kelas negatif memang benar-benar negatif. Secara keseluruhan, model ini menunjukkan kinerja yang sangat baik dengan akurasi

mencapai 98%. Hasil ini mengindikasikan bahwa model mampu mengklasifikasikan data dengan tingkat akurasi yang tinggi terdapat pada Gambar 10.

Classification report:

	precision	recall	f1-score	support
Negatif	0.85	0.99	0.91	490
Netral	0.98	0.83	0.90	438
Positif	0.99	0.93	0.96	283
accuracy			0.92	1211
macro avg	0.94	0.92	0.92	1211
weighted avg	0.93	0.92	0.92	1211

Gambar 10 Hasil Evaluasi Naive Bayes

Gambar 10 ini memberikan gambaran rinci tentang kinerja model dalam memprediksi sentimen teks menjadi tiga kategori: negatif, netral, dan positif. Dari gambar di atas, kita dapat melihat bahwa model memiliki akurasi keseluruhan sebesar 92%, yang berarti model dapat memprediksi dengan benar sekitar 92% dari seluruh data. Selain itu, metrik seperti precision, recall, dan F1-score memberikan informasi lebih spesifik untuk setiap kelas. Misalnya, model sangat baik dalam mengidentifikasi kelas negatif (precision 0.85), tetapi sedikit kurang baik dalam mengidentifikasi kelas netral (recall 0.83). Secara keseluruhan, model ini menunjukkan kinerja yang cukup baik.

8. WordCloud

WordCloud adalah visualisasi data berbentuk awan kata yang menampilkan frekuensi kemunculan kata dalam sebuah teks. Kata-kata yang sering muncul akan ditampilkan dengan ukuran font yang lebih besar, sedangkan kata yang jarang muncul akan lebih kecil. WordCloud berguna untuk melihat kata-kata kunci atau topik utama dalam suatu teks dengan cepat dan mudah.

a. WordCloud Kata Positif

Gambar 11 Sentimen Positif

WordCloud "Kata Positif" ini menampilkan visualisasi kata-kata yang sering muncul dan dianggap memiliki konotasi positif dalam sebuah kumpulan teks. Kata-kata yang berukuran lebih besar menunjukkan frekuensi kemunculan yang lebih tinggi, sehingga dapat diinterpretasikan sebagai kata-kata yang paling sering dikaitkan dengan sentimen positif. Dari WordCloud ini, kita dapat melihat bahwa kata-kata seperti "baik", "bagus", "senang", dan "suka" kemungkinan besar tidak muncul secara dominan, namun kata-kata seperti "bangun", "maju", dan "berkembang" mungkin memiliki konotasi positif dalam konteks teks yang dianalisis.

b. WordCloud Kata Negatif

Gambar 12 Sentimen Negatif

WordCloud "Kata Negatif" ini menampilkan visualisasi kata-kata yang sering muncul dan cenderung memiliki konotasi negatif dalam sebuah kumpulan teks. Kata-kata yang berukuran lebih besar menunjukkan frekuensi kemunculan yang lebih tinggi, sehingga dapat diinterpretasikan sebagai kata-kata yang paling sering dikaitkan dengan sentimen negatif. Dari WordCloud ini, kita dapat melihat kata-kata seperti "tidak", "jangan", "buruk", dan "jelek" kemungkinan besar sering muncul, yang mengindikasikan adanya sentimen negatif dalam teks.

c. WordCloud Kata Netral

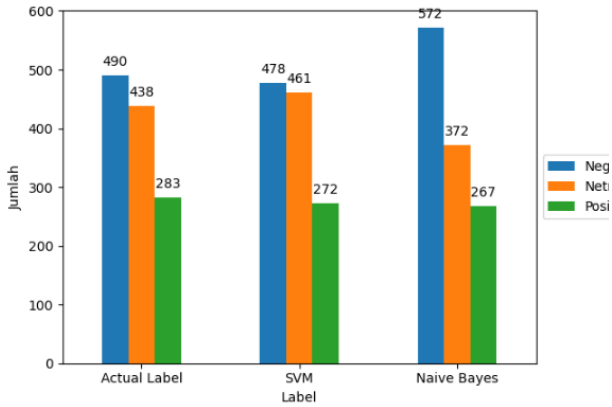
Gambar 13 Sentimen Netral

WordCloud "Kata Netral" ini menampilkan visualisasi kata-kata yang sering muncul dan cenderung tidak memiliki konotasi positif maupun negatif yang kuat dalam sebuah kumpulan teks. Kata-kata yang berukuran lebih besar menunjukkan frekuensi kemunculan yang lebih tinggi, sehingga dapat diinterpretasikan sebagai kata-kata yang paling sering digunakan dalam teks tanpa memberikan indikasi sentimen yang jelas. Dari

WordCloud ini, kita dapat melihat kata-kata seperti "yang", "ini", "saya", dan "kita" kemungkinan besar sering muncul, yang merupakan kata-kata penghubung umum dalam kalimat dan tidak memiliki muatan emosional yang kuat.

9. Diagram Batang

Diagram batang dipakai untuk membandingkan respons/komentar dengan frasa sentiment positif dan negatif, dan netral. Diagram batang sangat berguna untuk membandingkan data secara visual, terutama ketika kita ingin melihat perbedaan antara beberapa kategori atau kelompok. Dengan menggunakan diagram batang, kita dapat dengan mudah mengidentifikasi kategori mana yang memiliki nilai tertinggi atau terendah tersaji pada Gambar 14.



Gambar 14 Diagram Bar

Gambar 14 menampilkan histogram dari total 1211 data komentar tiktok menunjukkan bahwa persentase opini masyarakat untuk kelas pada sentimen positif sebesar 283 komentar data aktual label, 272 komentar data SVM label, dan 267 komentar data Naïve Bayes label, sedangkan untuk sentimen negatif sebesar 490 komentar data aktual label, 478 komentar data SVM label, dan 572 komentar data Naïve Bayes label, terakhir untuk sentimen netral sebesar 438 komentar data aktual, 461 komentar data SVM label, dan 372 komentar data Naïve Bayes label. Jadi hasil opini masyarakat di media sosial tiktok berada pada sentimen NEGATIF tersaji pada Tabel 2.

Table 1 Hasil Perbandingan 2 Metode

Metode	*Accuracy*
Support Vector Machine	97%
Naïve Bayes	92%

Dapat dilihat dari tabel 1 hasil analisis sentimen menggunakan Support Vector Machine dan Naïve Bayes dengan melakukan perhitungan bobot setiap kata pada opini masyarakat Tiktok terhadap topik Ibu Kota Nusantara. Dari total 1211 data komentar tiktok menunjukkan bahwa persentase perbandingan antara Metode Support Vector Machine dengan Naive Bayes berdasarkan tingkat akurasi yang diperoleh oleh masing-masing metode. Support Vector Machine memperoleh tingkat akurasi 98%, di mana tingkat akurasinya lebih tinggi daripada tingkat akurasi metode naive bayes dengan persentase 92%. Sehingga dapat dikatakan metode Support Vector Machine dalam melakukan analisis sentimen terhadap opini masyarakat mengenai ibu kota nusantara lebih baik dari pada metode Naïve Bayes.

4. Kesimpulan

Hasil penelitian ini menyimpulkan bahwa metode Support Vector Machine dan Naïve Bayes yang telah diimplementasikan dengan Bahasa pemrograman Python berhasil diterapkan untuk menganalisis sentimen masyarakat di Tiktok terhadap Ibu Kota Nusantara dengan mengklasifikasikan sebanyak 1529 data komentar ke kelas positif, negatif, dan netral. Teknik pembobotan TF-IDF digunakan untuk meningkatkan akurasi analisis. Hasilnya, sentimen masyarakat di Tiktok cenderung komentar kurang setuju. Metode Support Vector Machine dan Naïve Bayes terbukti cukup akurat dalam mengklasifikasikan sentimen masyarakat di Tiktok. Hal ini terlihat dari tingkat akurasi klasifikasi yang dihasilkan, yaitu 98% untuk metode Support Vector Machine dan 92% untuk metode Naïve Bayes. Support Vector Machine memiliki performa lebih baik dibandingkan Naïve Bayes dengan komentar yang berlabel “Negatif”. Kata – kata populer yang muncul dalam topik sentimen tentang Ibu Kota Nusantara di Tiktok antara lain IKN, Jakarta, kota, rakyat, indonesia, dll, yang tercermin dalam wordcloud dari hasil analisis sentimen.

Daftar Rujukan

[1] Q. A. A’yuniyah and M. Reza, “Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Di Sma Negeri 15 Pekanbaru,” *Indones. J. Inform. Res. Softw. Eng.*, vol. 3, no. 1, pp. 39–45, 2023, doi: 10.57152/ijirse.v3i1.484.

[2] D. Oktavia, Y. R. Ramadahan, and M. Minarto, “Analisis Sentimen Terhadap Penerapan Sistem E-Tilang Pada Media Sosial Twitter Menggunakan Algoritma Support Vector Machine (SVM),” *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 1, pp. 407–417, 2023, doi: 10.30865/klik.v4i1.1040.

[3] M. A. Djamaludin, A. Triayudi, and E. Mardiani, “Analisis Sentimen Tweet KRI Nanggala 402 di Twitter menggunakan Metode Naïve Bayes Classifier,” *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 6, no. 2, pp. 161–166, 2022, doi: 10.35870/jtik.v6i2.398.

[4] J. Florensius Sianipar, Y. R. Ramadhan, and I. Jaelani, “Analisis Sentimen Pembangunan Kereta Cepat Jakarta-Bandung di Media Sosial Twitter Menggunakan Metode Naive Bayes,” *Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 1, pp. 360–367, 2023, doi: 10.30865/klik.v4i1.1033.

[5] N. R. O. S. A. R. Lestari, “Implementation Of Text Mining And Pattern Discovery With Naive Bayes Algorithm For Classification Of Text Documents,” *Jzurnal Teknol. Inf. Komun. Digit. Zo.*, vol. 14, no. 1, pp. 88–102, 2023.

[6] H. Paul, A. Sartika Wiguna, and H. Santoso, “Penerapan Algoritma Support Vector Machine Dan Naive Bayes Untuk Klasifikasi Jenis Mobil Terlaris Berdasarkan Produksi Di Indonesia,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 39–44, 2023, doi: 10.36040/jati.v7i1.5555.

[7] E. Suryati, Styawati, and A. A. Aldino, “Analisis Sentimen Transportasi Online Menggunakan Ekstraksi Fitur Model

Word2vec Text Embedding Dan Algoritma Support Vector Machine (SVM)," *J. Teknol. Dan Sist. Inf.*, vol. 4, no. 1, pp. 96–106, 2023, [Online]. Available: https://doi.org/10.33365/jtsi.v4i1.2445

[8] F. Amaliah and I. K. Dwi Nuryana, "Perbandingan Akurasi Metode Lexicon Based Dan Naive Bayes Classifier Pada Analisis Sentimen Pendapat Masyarakat Terhadap Aplikasi Investasi Pada Media Twitter," *J. Informatics Comput. Sci.*, vol. 3, no. 03, pp. 384–393, 2022, doi: 10.26740/jinacs.v3n03.p384-393.

[9] J. Muliawan and E. Dazki, "Sentiment Analysis of Indonesia'S Capital City Relocation Using Three Algorithms: Naïve Bayes, Knn, and Random Forest," *J. Tek. Inform.*, vol. 4, no. 5, pp. 1227–1236, 2023, doi: 10.52436/1.jutif.2023.4.5.1436.

[10] S. Sumayah, F. Sembiring, and W. Jatmiko, "Analysis of Sentiment of Indonesian Community on Metaverse Using Support Vector Machine Algorithm," *J. Tek. Inform.*, vol. 4, no. 1, pp. 143–150, 2023, doi: 10.52436/1.jutif.2023.4.1.417.

[11] D. Pramana, M. Afdal, M. Mustakim, and I. Permana, "Analisis Sentimen Terhadap Pemindahan Ibu Kota Negara Menggunakan Algoritma Naive Bayes Classifier dan K-Nearest Neightbors," *J. Media Inform. Budidarma*, vol. 7, no. 3, pp. 1306–1314, 2023, doi: 10.30865/mib.v7i3.6523.

[12] P. Subarkah, P. W. Rahayu, I. Darmayanti, and R. Riyanto, "Sentiment Analysis on Reviews of Women'S Tops on Shopee Marketplace Using Naive Bayes Algorithm," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 9, no. 1, pp. 126–133, 2023, doi: 10.33480/jitk.v9i1.4179.

[13] P. Elisa and A. Rahman Isnain, "Comparison of Random Forest, Support Vector Machine and Naive Bayes Algorithms to Analyze Sentiment Towards Mental Health Stigma," *J. Tek. Inform.*, vol. 5, no. 1, pp. 321–329, 2024, [Online]. Available: https://doi.org/10.52436/1.jutif.2024.5.1.1817

[14] F. M. Sarimole and W. Septian, "Analisis Sentimen Masyarakat Terhadap Isu Penundaan Pemilu 2024 Pada Twitter Dengan Metode Naive Bayes Dan Support Vector Machine," *J. Sains dan Teknol.*, vol. 5, no. 3, pp. 890–899, 2024, [Online]. Available: http://ejournal.sisfokomtek.org/index.php/saintek/article/view/1359

[15] R. B. Dahlian and D. Sitanggang, "Sentiment Analysis of Digital Television Migration on Twitter Using Naïve Bayes Multinomial Comparison, Support Vector Machines, and Logistic Regression Algorithms," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 12, no. 2, pp. 280–288, 2023, doi: 10.32736/sisfokom.v12i2.1668.