3rd Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)
30 Agustus 2023 – Jakarta, Indonesia
Volume 2, Nomor 2, September 2023 - ISSN 2962-8628 (*online*)

IMPLEMENTASI NAÏVE BAYES CLASSIFIER TERKAIT PENCALONAN GANJAR PRANOWO SEBAGAI CALON PRESIDEN 2024 DI TWITTER

Fadila Salsabila[1*], Utomo Budiyanto[2]

[1,2]Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia

Email: [1*]fadilaasalsabila20@gmail.com, [2]utomo.budiyanto@budiluhur.ac.id
(* : corresponding author)

Abstrak- Seiring dengan perkembangan zaman dan teknologi saat ini, media sosial merupakan salah satu wadah untuk saling berkomunikasi, berbagi informasi dan berpendapat. *Twitter* merupakan jejaring sosial yang banyak digunakan dikalangan masyarakat Indonesia saat ini. Pada tahun 2024, Indonesia akan menghadapi Pemilihan Presiden (Pilpres) yang penting bagi arah dan masa depan negara. Salah satu Calon Presiden (Capres) pada Pilpres 2024 adalah Ganjar Pranowo. *Twitter* banyak digunakan oleh masyarakat sebagai platform untuk berdiskusi dan berpendapat mengenai berbagai topik, yang dimana saat ini lagi ramai tentang pemilihan calon presiden untuk 2024 mendatang. Masalah yang diangkat dalam penelitian ini yaitu bagaimana mengkalisifikasikan sentimen pada data *tweet*, mengetahui opini masyarakat di *Twitter* terkait dengan pencalonan Ganjar Pranowo sebagai calon presiden pada tahun 2024 berupa opini postif, negatif maupun netral, serta menghitung seberapa akurat metode *Naïve Bayes Classifier* dalam melakukan klasifikasi sentimen. Analisis sentimen terhadap opini masyarakat terhadap sosok Ganjar sebagai Calon Presiden 2024 di *Twitter* dilakukan untuk mengatasi masalah tersebut. Oleh karena itu, tujuan dari penelitian ini yaitu analisis sentimen masyarakat di *Twitter* terkait pencalonan Ganjar Pranowo sebagai Calon Presiden 2024. Penelitian ini menggunakan metode *preprocessing* berupa *case folding*, *cleaning*, *slangword*, *stopword*, dan *stemming*. Metode yang digunakan adalah *Naïve Bayes Classifier* untuk mengklasifikasikan data ke dalam kategori yang tepat, dimana data akan dibagi menjadi tiga sentimen yaitu positif, negatif dan netral. Dan *Bag of Words* digunakan sebagai metode untuk ekstraksi fitur. Hasil dari proses klasifikasi lebih dominan label positif dengan data yang digunakan sebanyak 300 data dengan rasio pembagian data 70% (210 data) untuk data latih dan 30% (90 data) buat data uji. Lalu dilakukan pengujian sehingga menghasilkan *accuracy* 75.56%, *precision* 75.56% dan *recall* 100.00%.

Kata Kunci: analisis sentimen, *bag of words*, *naïve bayes classifier*

IMPLEMENTATION OF NAÏVE BAYES CLASSIFIER RELATED TO THE NOMINATION OF GANJAR PRANOWO AS PRESIDENTIAL CANDIDATE FOR 2024 ON TWITTER

Abstract- *Along with the current development of technology and time, social media has become one of the platforms for communication, sharing information, and expressing opinions. Twitter is a social networking site that is widely used by Indonesian society today. In 2024, Indonesia will face an important Presidential Election (Pilpres) that will determine the direction and future of the country. One of the presidential candidates (Capres) in the 2024 Pilpres is Ganjar Pranowo. Twitter is widely used by the public as a platform to discuss and express opinions on various topics, especially the selection of presidential candidates. The problem addressed in this study is how to classify sentiments in tweet data, understand public opinions on Twitter regarding Ganjar Pranowo's candidacy as president in 2024 in the form of positive, negative, and neutral opinions, and determine the accuracy of the Naïve Bayes Classifier method in sentiment classification. Sentiment analysis of public opinions on Ganjar as a presidential candidate in 2024 on Twitter is conducted to address these issues. Therefore, the purpose of this study is to analyze public sentiment on Twitter regarding Ganjar Pranowo's candidacy as president in 2024. This study uses preprocessing methods such as case folding, cleaning, slangword, stopword, and stemming. The method used is Naïve Bayes Classifier to classify data into the right categories, where data will be divided into three sentiments: positive, negative, and neutral, and Bag of Words is used as a method for feature extraction. The classification process resulted in a dominant positive label with the usage of 300 data with a data distribution ratio of 70% (210 data) for training data and 30% (90 data) for testing data. Testing was then carried out, resulting in an accuracy of 75.56%, precision of 75.56%, and recall of 100.00%.*

Keywords*: sentiment analysis, bag of words, naïve bayes classifier*

1. PENDAHULUAN

Dalam kehidupan bermasyarakaat, mengungkapkan sentimen / pendapat kepada orang lain telah menjadi suatu aktivitas setiap harinya. Dalam proses demokrasi modern, media sosial seperti *Twitter* telah menjadi platform penting bagi masyarakat untuk berbagi pendapat termasuk dalam politik. Orang-orang sekarang dapat secara aktif berbagi pendapat mereka tentang tokoh-tokoh politik dan topik-topik yang relevan melalui *Twitter*. Postingan *Twitter* biasanya digunakan oleh pengguna untuk mengunggah informasi tentang diri mereka sendiri, berbagi informasi, dan menyebarkan berita. Konten postingan juga dapat menyampaikan emosi pengguna. Rumusan masalah pada penelitian ini adalah bagaimana klasifikasi sentimen dilakukan dengan menggunakan metode *Naïve Bayes Classifier* pada data tweet terkait dengan pencalonan Ganjar Pranowo sebagai calon presiden pada tahun 2024?, bagaimana opini dan sentimen masyarakat di *Twitter* terhadap pencalonan Ganjar Pranowo sebagai calon presiden pada tahun 2024?, dan seberapa akurat metode *Naïve Bayes Classifier* dalam melakukan klasifikasi sentimen terhadap data *tweet* terkait dengan pencalonan Ganjar Pranowo sebagai calon presiden pada tahun 2024?. Berdasarkan dengan rumusan masalah yang ada, peneliti akan membangun sebuah sistem analisis sentimen yang menggunakan metode *Naïve Bayes Classifier* untuk melakukan klasifikasi sentimen terhadap data *tweet* terkait dengan pencalonan Ganjar Pranowo sebagai calon presiden pada tahun 2024. Dataset akan diambil dari *Twitter* dengan mengumpulkan *tweet* dengan kata kunci "Ganjar Pranowo capres 2024", dengan jumlah data sebanyak 300 *tweet*. Dataset akan dilabeli secara manual menjadi tiga kategori yaitu sentimen positif, negatif, dan netral. Sebelum melakukan klasifikasi sentimen, data akan dipreprocess dan dilakukan *feature selection* untuk meningkatkan akurasi model. Adapun tujuan dari penelitian ini adalah untuk membangun sebuah sistem analisis sentimen yang efektif menggunakan metode *Naïve Bayes Classifier* untuk mengukur opini dan sentimen masyarakat di *Twitter* terkait dengan pencalonan Ganjar Pranowo sebagai calon presiden pada tahun 2024 dan melakukan analisis sentimen terhadap data tweet yang dikumpulkan dan mengimplementasikan rancangan model untuk mengukur tingkat akurasi dari metode *Naïve Bayes Classifier*. Dalam penelitian ini, kami berharap dapat memberikan kontribusi dalam memahami sentimen masyarakat di Twitter terhadap pencalonan Ganjar Pranowo sebagai Calon Presiden pada Pilpres 2024. Selain itu, diharapkan penelitian ini dapat membantu dalam melakukan analisis opini masyarakat secara lebih efisien dan akurat.

Menurut KBBI, sentimen adalah pandangan atau pendapat yang didasarkan pada perasaan yang berlebih-lebihan terhadap sesuatu. Media sosial seperti *Twitter, Facebook, Instagram*, dan *Youtube* merupakan tempat untuk menyampaikan opini. Opini tersebut dapat berupa ujaran kebencian atau puji-pujian yang dapat menimbulkan perdebatan di media social [1]. Analisis sentimen adalah proses komputasi yang banyak digunakan untuk memahami pendapat seseorang tentang suatu hal. Proses ini menggunakan teknik analisis teks pada data teks untuk memahami dan mengelompokkan emosi seseorang, baik positif maupun negatif. Faktor yang mempengaruhi penggunaan analisis sentimen adalah cara pengelolaan data teks yang berbeda-beda [2]. Analisis sentimen terdiri dari lima langkah, yaitu *crawling data, pre-processing, feature selection, classification*, dan *evaluation*. Analisis sentimen dapat mengubah data yang tidak terstruktur menjadi data yang terstruktur. Dengan adanya analisis sentimen, kita dapat mengevaluasi dan menghasilkan ide di berbagai bidang. Analisis sentimen juga dapat menganalisis peristiwa, pernyataan atau komentar yang kontroversial. [3]. Dalam analisis sentimen, *Twitter* sering dijadikan sumber data karena strukturnya yang mudah untuk dianalisis. Peneliti menggunakan ulasan berbahasa Indonesia di *Twitter* sebagai sumber data untuk analisis sentimen [4]. *Text mining* adalah sebuah proses yang dapat dilakukan untuk mengeksplorasi dan menganalisis sejumlah besar data teks yang tidak terstruktur. Proses ini memanfaatkan perangkat lunak yang dapat mengidentifikasi konsep, pola, topik, kata kunci, dan atribut lainnya dalam data. Dalam beberapa pandangan, text mining juga dikenal sebagai analisis teks. Fungsi dari analisis teks adalah untuk dapat memilah-milah set data dengan menggunakan teknik *text mining* untuk memperoleh data yang terstruktur dan lebih mudah dianalisis [5]. *Text Preprocessing* merupakan salah satu tahap dalam *Text Mining* yang berfokus pada pembersihan atau eliminasi segala *noise* pada teks. Tujuannya adalah untuk mencegah data yang tidak lengkap, gangguan dalam data, serta data yang tidak konsisten [6]. *Naïve Bayes Classifier* adalah metode klasifikasi yang menggunakan teorema Bayes dengan probabilitas sederhana dan independensi karakter yang tinggi. Metode ini cocok digunakan pada banyak jenis dataset, memiliki performa cepat dalam mengklasifikasi data, serta memiliki tingkat akurasi yang tinggi [7]. Ekstraksi fitur menggunakan metode Bag of Words untuk menghitung frekuensi kemunculan kata dalam data. Metode ini tidak memperhatikan urutan kata dan tata bahasa, namun tetap memperhatikan keragaman kata. Hasil perhitungan ini akan digunakan pada metode *Multinomial Naïve Bayes* dalam mewakili dokumen atau kalimat yang akan dijalankan analisis sentimen [8].

Pada penelitian sebelumnya telah banyak dilakukan terkait analisis sentimen menggunakan algoritma *Naïve Bayes Classifier* seperti, Penelitian ini mengembangkan sebuah aplikasi untuk menganalisis opini masyarakat terhadap layanan Grab dan Gojek menggunakan text mining dan algoritma *Naïve Bayes Classifier* untuk

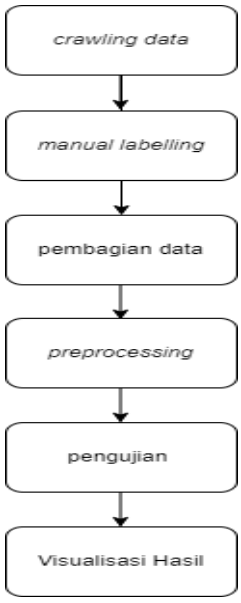
mengklasifikasikan tweet ke dalam kelas sentimen positif atau negatif. Data diperoleh melalui Twitter dengan kata kunci "grab" dan "gojek". Hasil akurasi klasifikasi data uji pada Grab dan Gojek berturut-turut adalah 74,34% dan 68,84%. Aplikasi ini juga menggunakan *Principal Component Analysis* (PCA) untuk menentukan faktor dari setiap sentimen yang telah divalidasi [9]. Penelitian ini menggunakan metode *naïve bayes classifier* dan *support vector machine* dalam mengekstraksi hasil. Tujuannya adalah untuk membandingkan tingkat keakurasiannya. Hasil penelitian menunjukkan bahwa algoritma *support vector machine* memiliki tingkat keakurasiannya sebesar 73,96%. Sedangkan untuk algoritma *naïve bayes classifier*, tingkat keakurasian yang diperoleh adalah sebesar 71,94%. Kedua algoritma tersebut diuji menggunakan dataset yang sama [10]. Penelitian ini menyajikan pemaparan terstruktur mengenai proses dan hasil implementasi NBC serta pengujian performanya dengan menggunakan confusion matrix. Dalam penelitian ini, diperoleh hasil akurasi sebesar 82%, presisi sebesar 93%, dan *recall* sebesar 52% [11]. Penelitian ini mengevaluasi sentimen masyarakat terhadap aspirasi yang disampaikan melalui *Twitter* dengan memperluas sistem yang sudah ada dan menggunakan metode klasifikasi *Naïve Bayes Classifier*. Data *tweet* diperoleh dari akun Twitter dengan kata kunci seperti #coronavirusindonesia atau #covid-19, dengan jumlah data tidak lebih dari 500 tweet. Setiap tweet akan melewati proses preprocessing, kemudian diklasifikasikan menjadi sentimen positif atau negatif. Dari hasil pengujian yang dilakukan pada 75 tweet, ditemukan bahwa akurasi *recall* adalah 32%, *precision* 80%, *F-Measure* 45%, dan rata-rata akurasi adalah 36% [12]. Penelitian ini menggunakan metode *Naïve Bayes Classifier* (NBC) dan pembobotan tf-idf untuk menentukan kelas sentimen dari tweet tentang toko JD.id. Penelitian juga menambahkan fitur konversi ikon emosi untuk meningkatkan akurasi klasifikasi. Hasil penelitian menunjukkan bahwa metode *Naïve Bayes* tanpa fitur tambahan mampu mengklasifikasikan sentimen dengan akurasi 96,44%. Namun, dengan penambahan fitur tf-idf dan konversi ikon emosi, nilai akurasi meningkat menjadi 98% [13]. Dari pengujian analisis sentimen menggunakan algoritma *Naïve Bayes* dengan ekstraksi fitur *Bag of Words* digunakan dan menghasilkan akurasi 89%, precision 83%, dan recall 87% [14]. Penelitian ini dilakukan untuk menganalisis sentimen pelanggan terhadap produk Shopee dengan menggunakan algoritma *Naïve Bayes Classifier*. Selain itu, penelitian ini juga menerapkan metodologi *Knowledge Discovery in Text* (KDT) untuk menggali informasi dari suatu data teks. Hasil klasifikasi menggunakan algoritma ini menunjukkan bahwa akurasi yang diperoleh mencapai 85% [15].

Penelitian ini menggunakan metode Bag of Words untuk melakukan ekstraksi fitur pada teks yang berupa tweet dan kemudian menerapkan algoritma Naive Bayes Classifier untuk mengklasifikasikan sentimen pada data tweet menjadi tiga kategori, yaitu positif, negatif, dan netral. Dengan menggunakan kedua metode tersebut, kami berhasil mengklasifikasikan sentimen pada data tweet terkait pencalonan Ganjar Pranowo sebagai Calon Presiden pada Pilpres 2024 dengan tingkat akurasi yang cukup baik. Opini publik penting dalam pemilihan presiden sehingga analisis sentimen diharapkan membantu memahami pandangan publik terhadap Ganjar Pranowo. Penulis berharap hasil dari penelitian ini dapat memberikan kontribusi pada pengembangan metode analisis sentimen pada data sosial media menggunakan metode Bag of Words dan algoritma Naive Bayes Classifier, serta dapat membantu para penggiat politik dan pemerintah untuk memahami sentimen masyarakat terkait dengan suatu topik tertentu.

2. METODE PENELITIAN

2.1 Penerapan Metode

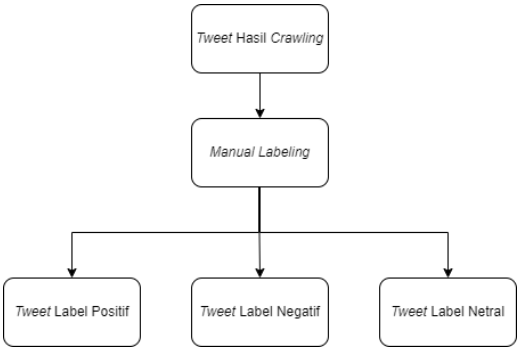
Penelitian ini menerapkan sistem menggunakan metode *Naïve Bayes* pada analisis sentimen di *Twitter*, terdapat rancangan sistem dari awal hingga akhir. Berikut merupakan tahapan-tahapan perancangan sistem yang digunakan dalam proses pembuatan aplikasi sesuai pada Gambar 1.



Gambar 1. Penerapan Metode

2.2 *Manual Labeling*

Pada tahapan *manual labeling*, data peneniliti divalidasi oleh dosen dari jurusan kriminologi Universitas Budi Luhur untuk melabelkan data. Pada proses ini pakar akan menentukan apakah data tersebut bernilai positif, negatif atau netral. Penjelasan tahap *manual labeling* dapat dilihat pada Gambar 2.



Gambar 2. *Manual Labeling*

2.3 *Naïve Bayes Classifier*

Naïve Bayes Classifier merupakan sebuah metoda klasifikasi yang berakar pada teorema Bayes. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yg dikemukakan oleh ilmuwan Inggris *Thomas Bayes*, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai *Teorema Bayes*. Ciri utama dari *Naïve Bayes Classifier* ini adalah asumsi yg sangat kuat (naïve) akan independensi dari masing-masing kondisi atau kejadian. Algoritme *Naive Bayes Classifier* masih sering memberikan hasil yang baik.

Pada algoritme *Naïve Bayes Classifier* atau bisa disebut sebagai Multinominal Naïve Bayes setiap dokumen di presentasikan dengan memasukkan "a1, a2, a3, …., an" dimana a merupakan kata pertama dan berikutnya sampai n, sedangkan V adalah label kategori, selanjutnya mencari nilai tertinggi dari kategori teks yang diujikan (V_{MAP}) persamaan (V_{MAP}) persamaan V_{MAP} dapat dilihat pada Persamaan (1).

$$V_{MAP} = \underset{vj \in V}{ar\ g\ max} P(v_j) \prod \mathrm{P}(a_i|v_j) \quad (1)$$

Nilai P(V_j) merupakan dihitung pada saat data latih. Nilai (V_j) didapatkan dari Persamaan (2).

$$P(V_j) = \frac{|Docs\ j|}{|training|} \quad (2)$$

Keterangan :
|Docs j| : Jumlah dokumen yang memiliki kategori j pada dokumen laith
|training| : Jumlah dokumen data latih
Setelah mendapatkan nilai (V_j), selanjutnya menentukan nilai $P(a_i|V_j)$ seperti pada Persamaan (3).

$$P(a_i|V_j) = \frac{|n_i + 1|}{|n + kosakata|} \quad (3)$$

Keterangan :
n_i : Jumlah kemunculan kata a_i pada dokumen laith yang berkategori v_j
n : Jumlah seluruh kata pada dokumen latih yang berkategori v_j
kosakata : Jumlah kata unik pada dokumen latih

2.4 *Confusion Matrix*

Confusion matrix adalah table yang digunakan untuk mengevaluasi kinerja suatu model klasifikasi dengan membandingkan prediksi model dengan nilai sebenarnya. *Confusion matrix* biasanya digunakan untuk masalah klasifikasi biner, dibana ada dua kelas yang diamati, yaitu kelas positif dan kelas negatif. Terdapat empat nilai yang dihasilkan *confusion matrix* yaitu *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN). Dapat dilihat pada Tabel 1.

Table 1. *Confusion Matrix*

	Positive	*Negative*
Positive	TP	FP
Negative	FN	TN

Keterangan :
True Positive (TP) : Jumlah data yang benar diprediksi sebagai positif
False Positive (FP): Jumlah data yang salah diprediksi sebagai positif
True Negative (TN) : Jumlah data yang benar diprediksi sebagai negative
False Negative (FN) : Jumlah data yang salah diprediksi sebagai negatif
Confusion matrix menghitung beberapa metrik evaluasi yang penting, antara lain :

a. Accuracy

Accuracy adalah mengukur nilai akurasi dari jumlah data yang benar diprediksi sebagai positif dan jumlah data yang benar diprediksi sebagai negatif dibagi jumlah seluruh data di database. Dapat dilihat pada Persamaan (4).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

b. Recall

Recall digunakan untuk menentukan seberapa baik model dalam mengidentifikasi peluang kasus dengan kategori positif yang benar diprediksi sebagai positif dibagi dengan keseluruhan data yang benar positif. Berikut pada Persamaan (5).

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

c. Precision

Precision adalah mengukur seberapa akurat model dalam mengklasifikasikan antara data yang diminta dengan hasil prediksi yang diberikan. Ini memberitahukan seberapa baik model dalam menghindari kesalahan dalam mengklasifikasikan data yang di prediksi positif dari semua kelas prediksi yang telah di prediksi dengan benar dan berapa banyak data yang benar-benar positif. Dapat dilihat pada Persamaan (6).

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

3. HASIL DAN PEMBAHASAN

3.1 *Crawling*

Data yang dikumpulkan berasal dari media sosial *Twitter*. Data yang dikumpulkan menggunakan *Google Collab* dengan kata kunci Ganjar Pranowo dan Capres 2024. Total data yang dikumpulkan sebanyak 500 data, kemudian dipilih 300 data yang sesuai untuk dijadikan dataset.

3.2 *Manual Labeling*

Pada penelitian ini, data dilabelisasi oleh dosen Fakultas Ilmu Sosial dan Studi Global Prodi Kriminologi. *Manual labeling* dilakukan untuk menentukan apakah data tersebut bernilai positif, negatif atau netral yang dapat dilihat pada Table 2 dan 3.

Table 2. Data Latih

Content	*Manual Labeling*
Dua bakal calon presiden (bacapres) Pilpres 2024, Ganjar Pranowo dan Anies Baswedan berbarengan menemui putra Raja Salman, Pangeran Mohammed Bin Salman (MBS). #Sindonews #news .https://t.co/EnyRkXrpSR	Positif
Sorak sorai dari Anak-Anak menyelimuti kala kehadiran Ganjar Pranowo di Jakarta. Para anak-anak tersebut berbaris untuk sekedar bersalaman dan bertatapan langsung oleh Sang Calon Presiden ini. #sahabatganjar #ganjarpranowo #ganjar #MEnangkanGAnjar #GenerasiGotongRoyong https://t.co/KeJHTDefE6	Negatif
Masyarakat Jakarta Percaya Bahwa Pak @ganjarpranowo Adalah Calon Presiden Terbaik Saat Ini Yang Cocok Untuk Penerus Kinerja Pakdhe @jokowi di 2024 Nanti. ???? #GanjarMenangTotal ?? https://t.co/U1UrxHJk9A	Netral
Elektabilitas Menteri BUMN Erick Thohir dalam bursa calon Wakil Presiden (Cawapres), bersaing dengan Menteri Pariwisata dan Ekonomi Kreatif Sandiaga Uno. Erick Thohir dinilai potensial mendongkrak kemenangan Ganjar Pranowo dalam Pilpres 2024. #GanjarErickDuetTerbaik https://t.co/lXrOHT51Am	Positif

Table 3. Sample Data Uji

Content	*Manual Labeling*
Bakal calon presiden PDI Perjuangan Ganjar Pranowo menyebut Presiden Jokowi selalu memantau gerak-gerik partai politik menjelang Pilpres 2024. #ganjarpranowo #jokowi #pilpres2024 #pemilu2024 #pdiperjuangan #gesuriid https://t.co/CIIwMJ3fLe	Positif

3.3 Pembagian Data

Tahapan pembagian data merupakan tahapan setelah *manual labeling* yang dimana data akan dibagi secara manual dengan perbandingan 70% untuk data latih dan 30% untuk data uji dari dataset. Seperti pada Table 4.

Table 4. Pembagian Data

Jenis Data	Jumlah
Dataset	300
Data Latih	210
Data Uji	90

3.4 *Preprocessing*

Tahap *preprocessing* merupakan tahapan untuk merbersihkan data dari kata-kata yang tidak penting dan tidak memiliki makna, sehingga menghasilkan data yang terstruktur. Proses ini dilakukan sesuai dengan hasil *crawling* data *Twitter*. Adapun tahapan dalam melakukan *preprocessing* sebagai berikut :

a. Case Folding

Case Folding adalah sebuah proses mengubah semua huruf menjadi huruf kecil. Dapat dilihat pada Table 5.

Table 5. Tahapan *Case Folding*

Data Asli	Tahapan *Case Folding*
Dua bakal calon presiden (bacapres) Pilpres 2024, Ganjar Pranowo dan Anies Baswedan berbarengan	dua bakal calon presiden (bacapres) pilpres 2024, ganjar pranowo dan anies baswedan berbarengan

menemui putra Raja Salman, Pangeran Mohammed Bin Salman (MBS). #Sindonews #news .https://t.co/EnyRkXrpSR	menemui putra raja salman, pangeran mohammed bin salman (mbs). #sindonews #news .https://t.co/enyrkxrpsr

b. Cleaning

Cleaning merupakan proses menghilangkan tanda baca, tag, hastag, link atau bisa disebut dengan proses pembersihan data. Dapat Dilihat dari Table 6.

Table 6. Tahapan *Cleaning*

Data Asli	Tahapan *Cleaning*
Dua bakal calon presiden (bacapres) Pilpres 2024, Ganjar Pranowo dan Anies Baswedan berbarengan menemui putra Raja Salman, Pangeran Mohammed Bin Salman (MBS). #Sindonews #news .https://t.co/EnyRkXrpSR	dua bakal calon presiden bacapres pilpres ganjar pranowo dan anies baswedan berbarengan menemui putra raja salman pangeran mohammed bin salman mbs

c. Slangword

Slangword merupakan tahapan untuk mengubah kata tidak baku menjadi kata baku. Dapat dilihat pada Table 7.

Table 7. Tahapan *Slangword*

Data Asli	Tahapan *Slangword*
Dua bakal calon presiden (bacapres) Pilpres 2024, Ganjar Pranowo dan Anies Baswedan berbarengan menemui putra Raja Salman, Pangeran Mohammed Bin Salman (MBS). #Sindonews #news .https://t.co/EnyRkXrpSR	dua bakal calon presiden bacapres pilpres ganjar pranowo dan anies baswedan berbarengan menemui putra raja salman pangeran mohammed bin salman mbs

d. Stopword

Stopword merupakan tahapan untuk menghilangkan kata tidak penting dalam sebuah data. Kata tidak penting pada tahapan *stopword* biasanya adalah kata penghubung seperti “dari”, “akan”, “atau”, “tapi”, “yang” dan lainnya. Tujuan utama dalam penerapan proses *Stopword* adalah mengurangi jumlah kata dalam sebuah dokumen. Dapat dilihat pada Table 8.

Table 8. Tahapan *Stopword*

Data Asli	Tahapan *Stopword*
Dua bakal calon presiden (bacapres) Pilpres 2024, Ganjar Pranowo dan Anies Baswedan berbarengan menemui putra Raja Salman, Pangeran Mohammed Bin Salman (MBS). #Sindonews #news .https://t.co/EnyRkXrpSR	calon presiden bacapres pilpres ganjar pranowo anies baswedan berbarengan menemui putra raja salman pangeran mohammed bin salman

e. Stemming

Stemming merupakan tahapan yang berfungsi untuk menghilangkan awalan, akhiran, imbuhan, atau kata depan dari suatu kata sehingga kata tersebut bentuknya kembali menjadi kata dasar. Tahap ini dilakukan agar kalimat yang digunakan untuk membuat *knowledge base*, serta data latih dan data uji hanya mengandung kata dasar saja sehingga akan memudahkan dalam pembuatannya. Dapat dilihat pada Table 9.

Table 9. Tahapan *Stemming*

Data Asli	Tahapan *Stemming*
Dua bakal calon presiden (bacapres) Pilpres 2024, Ganjar Pranowo dan Anies Baswedan berbarengan menemui putra Raja Salman, Pangeran Mohammed Bin Salman (MBS). #Sindonews #news .https://t.co/EnyRkXrpSR	calon presiden bacapres pilpres ganjar pranowo anies baswedan bareng temu putra raja salman pangeran mohammed bin salman

3.5 *Bag of Words*

Pada tahapan *Bag of Words* setelah tahapan *preprocessing* dengan menggabungkan semua data menjadi satu dan menghitung jumlah kata pada data latih. Lalu setelah itu data dipisahkan berdasarkan sentimen dan menghitung jumlah kemunculan kata uniknya seperti pada Table 10, 11 dan 12.

Table 10. Data Latih Positif

Content	Label
calon presiden bacapres pilpres ganjar pranowo anies baswedan bareng temu putra raja salman pangeran mohammed bin salman	Positif
elektabilitas menteri bumn erick thohir bursa calon wakil presiden cawapres saing menteri pariwisata ekonomi kreatif sandiaga uno erick thohir nilai potensial dongkrak menang ganjar pranowo pilpres	Positif

Table 11. Data Latih Negatif

Content	Label
sorak sorai anak limut hadir ganjar pranowo jakarta anak salam tatap langsung sang calon presiden	Negatif

Table 12. Data Latih Netral

Content	Label
masyarakat jakarta percaya pak calon presiden baik cocok terus kerja pakde	Netral

3.6 *Naïve Bayes Classifier*

Setalah melalui tahapan diatas, tahapan selanjutnya yaitu melakukan klasfikasi. Data yang digunakan yaitu data latih dan menghitung probabilitasnya.

Proses perhitungan probabilitas setiap sentimen :

a. Probabilitas sentimen Positif, dengan P adalah probabilitas :
$$p(positif) = \frac{2}{4} = 0{,}5$$
b. Probabilitas sentimen Negatif, dengan P adalah probabilitas:
$$p(negatif) = \frac{1}{4} = 0{,}25$$
c. Probabilitas sentimen Netral, dengan P adalah probabilitas :
$$p(netral) = \frac{1}{4} = 0{,}25$$

Jadi, hasil pengujian pada Algoritme *Naïve Bayes* adalah sentimen Positif memiliki probabilitas 0,5 sedangkan sentimen Negatif dan Netral memiliki probabilitas 0.25 dari 4 data latih yang dimana terdapat 2 data sentimen Positif dan 1 data untuk sentimen Negatif dan Netral.

Setelah menghitung nilai probabilitas setiap sentimen, selanjutnya yaitu menghitung jumlah kata unik pada setiap sentimen di data latih yang dimana hasil dari jumlah kata unik setiap sentimen tersebut akan dijumlahkan dan dijadikan sebagai total kosakata untuk menghitung probabililitas setiap kata menggunakan Persamaan (3) dan (2).

Proses pengujian data dilakukan setelah pelatiham model menggunakan data latih menggunakan data uji terhadap data latih. Proses ini dilakukan untuk membandingkan label asli dari data uji dan label yang di prediksi oleh *Naïve Bayes Classifier* untuk menghitung nilai performa seperti *accuracy*, *precision* dan *recall*.

3.7 Hasil Pengujian

3.1.1. Pengujian Klasifikasi

Pengujian klasifikasi menggunakan algoritme *Naïve Bayes Classifier* dilakukan untuk menghitung nilai probabilitas setiap sentimen yaitu positif, negatif dan netral. Sentimen yang memiliki nilai probabilitas tertinggi akan digunakan sebagai label untuk data yang diuji. Dapat dilihat pada Tabel 18.

Table 13. Pengujian Klasifikasi

No.	Data Uji	Label Actual	Label Predicted	Label Hasil
1.	bapak ganjar pranowo jadi calon presiden republik indonesia tahun	Positif	Positif	TP
2.	menang telak raih calon presiden bacapres gerindari a prabowo subianto head to head lawan bacapres pdip ganjar pranowo prabowo subianto	Netral	Positif	FP
3.	breaking news calon presiden pdi juang ganjar pranowo yakin milik empat faktor menetukan menang pilpres jokowi effect dukung partai politik figur cawapres soliditas tim menang ganjar pranowo	Positif	Positif	TP

4.	bak air dukung alir bapak ganjar pranowo calon presiden ri ibuibu majelis taklim muslimah kabupaten tanggung	Positif	Positif	TP
5.	ganjar pranowo calon presiden ide ide hebat maju sektor umkm	Negatif	Positif	FP
….90.	sukses dukung atas tetap ganjar pranowo calon presiden	Positif	Positif	TP

3.1.2. Pengujian *Confusion Matrix*

Pengujian model klasifikasi dilakukan dengan menggunakan *confusion matrix* untuk mengevaluasi kinerja model. *Confusion matrix* ini digunakan untuk membandingkan hasil klasifikasi berdasarkan empat kemungkinan hasil, yaitu *true positive, false positive, true negative, false negative, true neutral* dan *false neutral*. Dengan demikian, *confusion matrix* menjadi alat yang dapat memberikan gambaran seberapa baik performa model klasifikasi dalam mengklasifikasikan data. Dapat dilihat pada Tabel 19.

Table 14. Pengujian *Confusion Matrix*

TP	FN	FNet	FP	TN	FNet	FPNet	FNeg	Tnet
68	0	0	10	0	0	12	0	0

Proses perhitungan nilai *accuracy, precision* dan *recall* :

a. Perhitungan *accuracy* :

$$accuracy = \frac{TP + TN + TNet}{total\ keseluruhan}\ x\ 100\%$$
$$accuracy = \frac{68 + 0 + 0}{68 + 0 + 0 + 10 + 0 + 0 + 12 + 0 + 0}\ x\ 100\%$$
$$accuracy = \frac{68}{90} x\ 100\%$$
$$accuracy = 75.56\%$$

b. Perhitungan *precision* :

$$precision = \frac{TP}{TP + FP + FPNet}\ x\ 100\%$$
$$precision = \frac{68}{68 + 10 + 12}\ x\ 100\%$$
$$precision = \frac{68}{90}\ x\ 100\%$$
$$precision = 75.56\%$$

c. Perhitungan *recall* :

$$recall = \frac{TP}{TP + FN + FNet}\ x\ 100\%$$
$$recall = \frac{68}{68 + 0 + 0}\ x\ 100\%$$
$$recall = \frac{68}{68}\ x\ 100\%$$
$$recall = 100.00\%$$

4. KESIMPULAN

Dalam pembuatan aplikasi analisis sentimen menggunakan algoritme *Naïve Bayes Classifier* dengan beberapa tahapan seperti *crawling* data, *manual labeling*, pembagian data, *preprocessing*, ekstraksi fitur menggunakan *bag of words* dan evaluasi. Sumber dataset yang digunakan dari *Twitter* dengan tweet berbahasa Indonesia dengan kata kunci "Ganjar Pranowo capres 2024" dengan jumlah data yang digunakan sebanyak 300 data dari *Twitter*. Pelabelan dataset dari tweet dilakukan secara manual dibagi menjadi tiga kategori yaitu sentimen positif, negatif dan netral. Berdasarkan 300 data *tweet* yang digunakan dengan 70% (210 data) untuk data latih dan 30% (90 data) untuk data uji, pandangan atau sentimen masyarakat terkait pencalonan Ganjar Pranowo sebagai calon presiden 2024 bernilai positif. Hasil yang didapatkan dari penelitian ini yaitu *accuracy* 75.56%, *precision* 75.56% dan *recall* 100.00%. Adapun manfaat dari penelitian ini untuk masyarakat dapat diklasifikasikan dalam beberapa aspek.

Penelitian ini diharapkan dapat memberikan informasi yang akurat dan terpercaya mengenai pandangan masyarakat terhadap pencalonan Ganjar Pranowo sebagai Calon Presiden 2024. Penelitian ini memberikan gambaran kepada masyarakat mengenai kegunaan analisis sentimen dalam memahami opini publik terkait suatu topik atau tokoh politik.

Untuk penelitian selanjutnya, perlu dilakukan penelitian yang lebih luas pada berbagai sumber data dalam Bahasa Indonesia seperti media sosial, situs berita, forum diskusi dan lain-lain. Dalam melakukan *manual labeling* lebih baik melibatkan lebih banyak responden agar lebih variatif dan bisa mempresentasikan banyak pendapat. Untuk penelitian selanjutnya dapat menggunakan algoritme yang lain atau lebih dari satu algoritme untuk membandingkan akurasi dan kecepatan pemrosesan dalam pemodelan analisis sentimen.

DAFTAR PUSTAKA

[1] [1] D. Rusdiaman and D. Rosiyadi, "*Analisa Sentimen Terhadap Tokoh Publik Menggunakan Metode Naive Bayes Classifier dan Support Vector Machine*," vol. 4, issue 2, 2019.
[2] [2] W. Widayat, "*Analisis Sentimen Movie Review Menggunakan Word2Vec dan Metode LSTM Deep Learning*," Jurnal Media Informatika Budidarma, vol. 5, no. 3, pp. 1018-1026, 2021, doi: 10.30865/mib.v5i3.3111.
[3] [3] A. P. Natasuwarna Jurusan Sistem Informasi and S. Pontianak, "*Seleksi Fitur SVM pada Analisis Sentimen Keberlanjutan Pembelajaran Daring*," vol. 19, issue 4, 2020.
[4] [4] S. F. Handayani, R. W. Pratiwi, D. Dairoh, and D. I. Af'idah, "Analisis Sentimen pada Data Ulasan Twitter dengan Long-Short Term Memory," JTERA (Jurnal Teknologi Rekayasa), vol. 7, no. 1, pp. 39-46, 2022, doi: 10.31544/jtera.v7.i1.2022.39-46.
[5] [5] M. Mega, M. Olhang, S. Achmadi, and F. X. Ariwibisono, "*Analisis Sentimen Pengguna Twitter terhadap Covid-19 di Indonesia Menggunakan Metode Naive Bayes Classifier (NBC)*," in Jurnal Mahasiswa Teknik Informatika, vol. 4, issue 2, 2020.
[6] [6] F. V. Sari and A. Wibowo, "*Analisis Sentimen Pelanggan Toko Online jD.ID Menggunakan Metode Naive Bayes Classifier Berbasis Konversi Ikon Emosi*," Jurnal SIMETRIS, vol. 10, no. 2, 2019.
[7] [7] D. Normawati and S. A. Prayogi, "*Implementasi Naive Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter*," in Jurnal Sains Komputer & Informatika (J-SAKTI), vol. 5, issue 2, 2021.
[8] [8] V. R. Albahar Sejati and W. Pramusinto, "*2nd Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI) 21 Maret 2023-Jakarta*," vol. 2, no. 1, 2023.
[9] [9] Olive, D. Putra, K. Rega Prilianti, P. Lucky, dan T. Irawan, "*Implementasi Text Mining untuk Analisis Opini Masyarakat terhadap Kinerja Layanan Transportasi Online dengan Analisis Faktor*," vol. 8, no. 2, 2020.

[10] [10] D. Rusdiaman dan D. Rosiyadi, "*Analisa Sentimen Terhadap Tokoh Publik Menggunakan Metode Naïve Bayes Classifier dan Support Vector Machine*," vol. 4, no. 2, 2019.
[11] [11] D. Normawati dan S. A. Prayogi, "*Implementasi Naïve Bayes Classifier dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter*," J. Sains Komput. Informatika (J-SAKTI), vol. 5, no. 2, 2021.
[12] [12] M. Mega, M. Olhang, S. Achmadi, dan F. X. Ariwibisono, "*Analisis Sentimen Pengguna Twitter terhadap COVID-19 di Indonesia Menggunakan Metode Naïve Bayes Classifier (NBC)*," in *Jurnal Mahasiswa Teknik Informatika*, vol. 4, issue 2, 2020.
[13] [13] F. V. Sari dan A. Wibowo, "*Analisis Sentimen Pelanggan Toko Online JD.ID Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi*," *Jurnal SIMETRIS*, vol. 10, no. 2, 2019.

[14] [14] A. I. Tanggraeni dan M. N. N. Sitokdana, "*Analisis Sentimen Aplikasi E-Government pada Google Play Menggunakan Algoritma Naïve Bayes*," vol. 9, no. 2, hal. 785–795, 2022.
[15] [15] L. Oktaria Sihombing dan B. Arif Dermawan, "Sentimen Analisis Customer Review Produk Shopee Indonesia Menggunakan Algoritma Naïve Bayes Classifier," *Edumatic: Jurnal Pendidikan Informatika*, vol. 5, no. 2, hal. 233-242, 2021. [Online]. Tersedia: https://doi.org/10.29408/edumatic.v5i2.4089.