Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika dan Komputer)
Volume 23 ; Nomor 1 ; Februari 2024 ; Page 37-45
E-ISSN : 2615-3475; P-ISSN : 1978-6603
https://ojs.trigunadharma.ac.id/index.php/jis/index

Klasifikasi Emosi Pengguna Twitter Terhadap Bakal Calon Presiden Pada Pemilu 2024 Menggunakan Algoritma Naïve Bayes

Khusnul Arifin[1], Said Iskandar Al Idrus[2]

[1] Ilmu Komputer, Universitas Negeri Medan, Medan, Indonesia
[2]Ilmu Komputer, Universitas Negeri Medan, Medan, Indonesia
Email: [1]khusnul.arfn@gmail.com, [2]said_iskandar@gmail.com
Email Penulis Korespondensi: khusnul.arfn@gmail.com

Article History:
Received Dec 09th, 2023
Revised Dec 22th, 2023
Accepted Jan 12th, 2024

Abstrak

Tahun 2024 mendatang akan menjadi tahun politik. Terdapat tiga bakal calon presiden yang namanya telah muncul di masyarakat, yaitu Anies Baswedan, Ganjar Pranowo, dan Prabowo Subianto. Banyak berseliweran di media sosial terkhususnya Twitter respon mengenai para tokoh politik tersebut. Beragamnya respon para pengguna Twitter terhadap bakal calon presiden di pemilu 2024 mengakibatkan juga banyaknya jenis dari emosi "cuitan" para penggunanya, oleh karena itu diperlukan adanya analisis untuk mengetahui pandangan masyarakat terhadap para bakal calon presiden tersebut berdasarkan klasifikasi emosinya. Klasifikasi emosi pada proses penelitian ini menggunakan algoritma *Multinomial Naïve Bayes* dengan melibatkan beberapa tahapan proses, seperti *preprocessing data*, pelabelan data, ekstraksi fitur, pembagian dataset, klasifikasi, dan evaluasi model. Menggunakan pembagian data dengan metode 80:20, yaitu data latih dibagi menjadi 80% dan data uji menjadi 20%. Hasil pelabelan emosi dari tiap bakal calon presiden menunjukkan nilai yang berbeda satu sama lain. Tahapan pengujian dilakukan dan dibagi dengan beberapa tahap, yaitu pengujian menggunakan label netral, tanpa label netral, dan *random sampling*. Performa model algoritma Naïve Bayes tanpa menggunakan label netral menunjukkan performa yang lebih baik, dengan nilai akurasi model sebesar 58% pada data Anies Baswedan, 58% pada data Prabowo Subianto, dan 76% pada data Ganjar Pranowo, serta gabungan 69%. Klasifikasi pada skenario pengujian menggunakan label netral menunjukkan akurasi sebesar 55% pada data Anies Baswedan, 60% pada data Ganjar Pranowo, dan 53% pada data Prabowo Subianto, sedangkan untuk gabungan semuanya nilai akurasi sebesar 51%.
Kata Kunci : Naïve Bayes, Pemilu 2024, Emosi, Klasifikasi

Abstract

2024 will be a political year. There are three presidential candidates whose names have appeared in the public, namely Anies Baswedan, Ganjar Pranowo, and Prabowo Subianto. There are many responses circulating on social media, especially Twitter, regarding these political figures. The various responses of Twitter users to the presidential candidates in the 2024 election have resulted in many types of emotional "tweets" from their users, therefore analysis is needed to determine the public's views on the presidential candidates based on their emotional classification. Emotion classification in this research process uses the Multinomial Naïve Bayes algorithm which involves several process stages, such as data preprocessing, data labeling, feature extraction, dataset division, classification, and model evaluation. Using data division with the 80:20 method, namely training data is divided into 80% and test data into 20%. The results of labeling the emotions of each presidential candidate show different values from each other. The testing stages were carried out and divided into several stages, namely testing using neutral labels, without neutral labels, and random sampling. The performance of the Naïve Bayes algorithm model without using neutral labels shows better performance, with a model accuracy value of 58% on Anies Baswedan's data, 58% on Prabowo Subianto's data, and 76% on Ganjar Pranowo's data, and a combined 69%. Classification in the test scenario using neutral labels showed an accuracy of 55% on Anies Baswedan's data, 60% on Ganjar Pranowo's data, and 53% on Prabowo Subianto's data, while for all combined the accuracy value was 51%.

1. PENDAHULUAN

Media sosial di saat ini merupakan salah satu bagian penting dari kehidupan bersosial masyarakat. Hampir seluruh kalangan umur terkhususnya di Indonesia memiliki media sosial. Berdasarkan data yang dicatat oleh We Are Social, melansir bahwa total pengguna media sosial aktif yang ada di Indonesia sebanyak 167 juta orang pada Januari 2023, dan jika dibandingkan dengan jumlah populasi di Indonesia, angka tersebut setara dengan 60.4% dari populasi masyarakat Indonesia.

Twitter merupakan salah satu media sosial terkenal, dan banyak digunakan di Indonesia. Berdasarkan survei yang telah dilakukan oleh Global Web Index (GWI) pada kuartal 3 tahun 2022 berdasarkan persentase pengguna internet berusia 16-64 tahun, Twitter termasuk ke dalam peringkat ke-7 dengan pengguna terbanyak di Indonesia. Sebanyak 60.2% pemakai internet di Indonesia mengakses twitter, dengan jumlah sekitar 24 juta pengguna di Indonesia pada tahun 2023.

Twitter *platform* media sosial yang memberikan akses kepada para pengguna untuk membagikan pesan yang disebut dengan *tweet* [1]. Terdapat satu fitur yang andalan yang terpada pada Twitter, yaitu tweet, para pengguna diberikan kebebasan untuk mem-posting apapun, baik dalam bentuk teks, foto, bahkan video. Berdasarkan data dari Statista terdapat 500 juta tweet diterbitkan tiap harinya, dan 350.000 tweet di-posting setiap menit. Topik tweet yang ada pada Twitter juga beragam, mulai dari hobi, opini, dan juga kritik terhadap suatu hal. Twitter Indonesia menghimpun topik yang paling ramai dibicarakan pengguna di Indonesia, dan tercatat bahwa politik menjadi salah satu pembahasan yang sedang hangat dan sering dibahas di kalangan pengguna Twitter di Indonesia.

Pemilu 2024 menjadi ajang tahun politik di Indonesia. Mulai dari tahun ini, banyak berseliweran di media sosial mengenai politik, baik berupa kampanye ataupun isu mengenai politik itu sendiri. Berdasarkan hal tersebut topik politik juga semakin marak dibicarakan di Twitter oleh masyarakat Indonesia. Salah satu topik politik yang sering dibicarakan adalah mengenai calon presiden.

Beberapa bakal calon presiden berdasarkan hasil survei yang sudah dilakukan menghasilkan tiga nama calon presiden pada pemilu 2024 nanti, antara lain Ganjar Pranowo, Anies Baswedan, dan Prabowo Subianto. Nama-nama tokoh politik diatas nantinya akan dijadikan sebagai kata kunci dalam klasifikasi emosi dari para pengguna Twitter terhadap pilpres 2024, dan akan dijadikan acuan untuk melihat respon masyarakat.

Respon yang diberikan masyarakat Indonesia terhadap topik politik tersebut juga beragam, ada yang membenci dan juga mendukung. Tidak jarang adanya doa, ungkapan senang, dan bahkan ujaran kebencian ditemukan di Twitter mengenai bakal calon presiden yang terlibat dalam tahun politik. Beragamnya emosi dari masyarakat sangat erat kaitannya dengan tahun politik yang sedang berlangsung saat ini, bahkan fenomena ini sudah berlangsung dari tahun ke tahun.

Banyaknya konten dan juga respon dari masyarakat kepada suatu tokoh politik sering berdampak negatif terhadap media sosial di Indonesia, salah satunya yaitu munculnya konten hoax. Tidak jarang muncul konten-konten hoax dengan tujuan untuk menjatuhkan dan memberikan citra buruk terhadap suatu tokoh politik, dan hal ini sudah menjadi fenomena lumrah yang terjadi menjelang tahun politik. Hal yang sama juga terjadi pada tahun 2019 silam, di momen yang sama yaitu ketika tahun politik. Berdasarkan data dari Kementerian Komunikasi dan Informatika (Kominfo) pada tahun 2019 lalu, jumlah konten hoax yang teridentifikasi pada tahun 2018 – 2019 silam sebanyak 1.731 konten. Konten hoax yang memicu munculnya ujaran kebencian diidentifikasi sebanyak 486 unggahan selama bulan April 2019.

Beragamnya respon para pengguna Twitter terhadap bakal calon presiden di pemilu 2024 mengakibatkan juga banyaknya jenis dari emosi "cuitan" para penggunanya. Emosi sendiri merupakan reaksi yang diberikan manusia sebagai respons terhadap suatu kondisi atau peristiwa. Reaksi yang diberikan melibatkan perubahan-perubahan fisik dan psikologi terhadap rangsangan yang diterima. Emosi juga dapat diartikan sebagai perasaan intens yang muncul terhadap seseorang atau sesuatu dan merupakan reaksi terhadap situasi tertentu. [1]

Teks adalah jenis media yang dipakai dalam berkomunikasi dan memberikan informasi. Selain sebagai penyampai informasi, teks juga digunakan untuk mengekspresikan emosi. Emosi yang diberikan manusia terbagi menjadi beberapa jenis, dalam penelitian ini, emosi yang akan digunakan dibagi menjadi ke dalam enam jenis emosi, yaitu emosi senang, sedih, marah, takut, terkejut, jijik, dan ditambah dengan satu emosi netral untuk data teks yang tidak memiliki emosi apapun. Keenam emosi ini ditemukan pertama kali oleh Paul Ekman, seorang psikolog yang menjadi perintis dalam studi emosi. Teori Paul Ekman telah banyak digunakan dalam bidang klasifikasi emosi, salah satu penelitian yang menerapkan teori ini adalah pengenalan emosi dari ekspresi wajah manusia [2].

Text mining sendiri menjadi variasi lain dari *data mining* yang mana tujuan dari metode ini untuk menemukan pola yang menarik dari sekumpulan data berbentuk teks yang jumlahnya banyak [3]. Pembeda dari dua hal ini adalah pola yang digunakan. *Text mining* mengambil dari sekumpulan data teks yang tidak terstruktur, sedangkan *data mining* biasanya sudah terdapat susunan data yang memiliki struktur. [4]

Klasifikasi emosi pada teks merupakan penerapan metode yang digunakan untuk dapat membagi jenis-jenis teks berdasarkan emosinya. Penerapan dari metode ini sendiri bertujuan untuk memahami perasaan yang disampaikan penulis, lalu juga untuk mengidentifikasi perasaan positif atau negatif seseorang terhadap suatu hal. Klasifikasi memiliki empat komponen, yaitu *class*, *predictors variable*, *training dataset*, *testing dataset* [5].

Algoritma Naïve Bayes adalah algoritma klasifikasi yang akan digunakan pada penelitian ini. Algoritma Naive Bayes memiliki kelebihan pada bentuk model yang sederhana, akan tetapi algoritma ini masih dapat bersaing dengan model algoritma yang lainnya. Pada penelitian terdahulu Naïve Bayes adalah salah satu pengklasifikasian paling sederhana dalam bidang machine learning, namun algoritma ini masih mendukung mesin vektor. Naïve Bayes sendiri juga memiliki kemampuan kecepatan yang efisien, karena hanya memerlukan waktu yang sedikit dalam melakukan pelatihan [6].

Algoritma klasifikasi sederhana yang disebut Naive Bayes Classifier terbukti efektif dan efisien dalam mengolah database dengan jumlah data yang besar, hasilnya diketahui cepat dan akurat.

Algoritma Naïve Bayes memiliki keunggulan yaitu cepat dihitung dan akurat meskipun sederhana. Untuk memastikan perkiraan yang dibutuhkan dalam proses klasifikasi, Naïve Bayes tidak memerlukan banyak data pelatihan (Data Training). Jalur perhitungan yang lebih pendek membuat algoritma ini lebih mudah digunakan. Berdasarkan penelitian yang sudah dilaksanakan oleh [7]. Metode Naïve Bayes menerapkan model *Confusion Matrix* dengan 210 jumlah data, 154 sebagai data training dan 56 data testing akurasi yang dihasilkan sebesar 82,14%. *Confusion matrix* merupakan Salah satu metode yang digunakan untuk mengevaluasi suatu metode klasifikasi adalah dengan melakukan perbandingan hasil klasifikasi sistem dengan klasifikasi aktual [4]. *Confusion matrix* biasanya mencakup informasi mengenai akurasi, *recall*, *precision*, dan *error ratio* [8].

Berdasarkan penelitian yang sudah dilaksanakan oleh Metode *Naïve Bayes* menerapkan model *Confusion Matrix* dengan 210 jumlah data, 154 sebagai *data training* dan 56 *data testing* akurasi yang dihasilkan sebesar 82,14%.

Metode Naïve Bayes dapat diterapkan pada pengklasifikasian data yang berupa berbentuk teks. Dalam penelitian yang dilakukan sebelumnya oleh [9], metode Multinominal Naïve Bayes dapat digunakan untuk pengolahan data, terutama dalam pengklasifikasian teks. Penelitian yang sama juga dilakukan oleh [10]. Metode Naïve Bayes lumrahnya sering digunakan dalam penelitian analisis sentimen. Biasanya pada penelitian analisis sentimen terdapat satu penggunaan metode penting praproses penghapusan kata (stopword removal) mendapatkan hasil yang efektif dalam proses pengklasifikasian analisis sentimen.

Python merupakan suatu bahasa pemrograman yang berorientasi pada interpretasi dan memiliki sifat yang sangat fleksibel dalam pemakaiannya. Bahasa ini menekankan pada kemudahan dalam membaca kode program dan diklaim sebagai bahasa yang efisien serta mudah dipelajari. Python dapat digunakan untuk mengembangkan berbagai jenis aplikasi, seperti aplikasi desktop, aplikasi web, dan aplikasi lainnya [11]. Python juga bahasa pemrograman *freeware* atau bisa disebut bebas digunakan, tidak terdapat batasan dalam penggunaan atau pendistribusiannya [12].

Dalam penelitian menggunakan teks, diperlukan adanya *pre-processing*, salah satunya Algoritma Nazief dan Adriani. Algoritma Nazief dan Adriani adalah suatu teknik *stemming* dalam bahasa Indonesia yang berfungsi untuk mengganti kata-kata yang berbentuk infleksi menjadi kata dasar[13]. Sastrawi adalah *library* pada bahasa pemrograman *Python* yang dibangun dengan menggunakan algoritma NA (Nazief dan Adriani). *Library* tersebut berguna untuk melakukan penyederhanaan kata-kata yang telah mengalami infleksi dalam bahasa Indonesia, serta dapat juga digunakan untuk mengubah kata berimbuhan menjadi bentuk kata dasar[14].

Hasil yang diharapkan dari penelitian ini adalah sebuah model yang dapat digunakan untuk klasifikasi emosi berdasarkan teks tweet para pengguna Twitter di Indonesia. Data emosi pengguna ini dapat digunakan sebagai acuan untuk mengidentifikasi emosi dan juga pendapat masyarakat Indonesia terhadap tokoh politik di pemilu 2024 nanti.

2. METODOLOGI PENELITIAN

2.1 Lokasi dan Waktu Penelitian

Lokasi pada penelitian ini akan dilaksanakan di laboratorium komputer Program Studi Ilmu Komputer Universitas Negeri Medan dengan tujuan mengumpulkan studi literatur dan juga data *tweet* dari media sosial Twitter mengenai bakal calon presiden di pemilu 2024.

2.2 Jenis Penelitian

Penelitian ini menggunakan metode penelitian *Sequential Exploratory Design* dengan pendekatan *machine learning*. Penelitian *Sequential Exploratory Design* merupakan metode gabungan antara kualitatif dan kuantitatif. Metode ini dimulai dengan penelitian kualitatif untuk memperoleh pemahaman yang lebih mendalam tentang fenomena yang diteliti, kemudian dilanjutkan dengan penelitian kuantitatif untuk menguji hipotesis yang telah dihasilkan dari penelitian kualitatif.

Pendekatan *machine learning* sendiri merupakan penelitian untuk melakukan inferensi terhadap data berdasarkan landasan ilmu matematika, statistika, dan komputasi menggunakan bahasa program *Python*.

2.3 Populasi dan Sampel

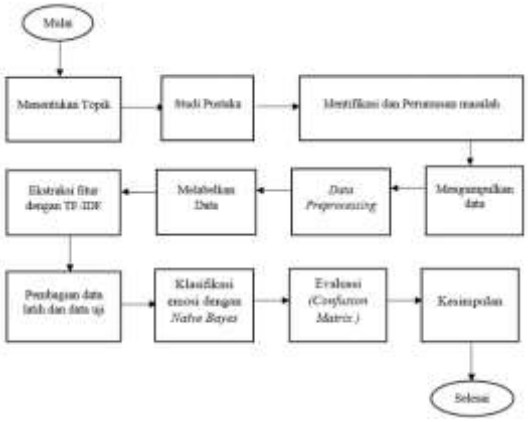
Populasi dalam penelitian ini adalah teks dari tweet yang diposting oleh pengguna di platform media sosial Twitter pada bulan Juli 2023, khususnya pada minggu pertama. Data ini diperoleh melalui proses crawling menggunakan layanan situs web pihak ketiga, Netlytic. Proses crawling dilakukan dengan mencari kata kunci yang terkait dengan nama para bakal calon presiden, yakni Anies Baswedan, Ganjar Pranowo, dan Prabowo Subianto, dengan pembagian 5.000 data tweet terhadap masing-masing calon. Terdapat 3 bakal calon presiden dalam pemilu 2024, sehingga total keseluruhan data adalah 15.000 data tweet. Menurut [15] ukuran sampel yang ideal adalah 10% dari populasi. Oleh sebab itu sampel yang akan digunakan adalah 10% dari total keseluruhan data tweet.

2.4 Variabel Penelitian

Variabel dalam penelitian akan diperiksa menjadi bentuk apa pun yang dipilih peneliti, dan setelah pengumpulan data, sebuah kesimpulan tercapai. Variabel pada penelitian ini adalah:

1. Kata Dasar Teks
   Merupakan variabel berbentuk kata dasar dari tweet berbahasa Indonesia setelah proses data pra-processing untuk dihitung frekuensinya dan menjadi kata kunci dalam proses perhitungan nantinya.
2. Bobot Kata
   Variabel ini mencakup bobot kata yang menunjukkan tingkat kepentingan kata dalam dokumen atau koleksi dokumen. Setiap kata memiliki bobot yang menggambarkan seberapa sering kata tersebut muncul dalam dokumen dan seberapa penting kata tersebut dalam konteks keseluruhan koleksi dokumen. Bobot kata ini dapat direpresentasikan ke dalam bentuk numerik dengan menggunakan matriks dimensi (dokumen x kata) atau biasa disebut dengan matriks TF-IDF.

2.5 Tahapan Penelitian



Gambar 1. Alur Tahapan Penelitian

1. Menentukan Topik
   Langkah utama yang dilakukan dalam penelitian ini adalah penentuan topik dari penelitian.
2. StudiPustaka
   Setelah topik ditentukan, dilanjutkan dengan studi pustaka mengenai metode yang akan digunakan dan berhubungan dengan penelitian. Pada tahap ini dilakukan proses mencari, mempelajari, dan memahami mengenai apa saja yang bersangkutan dengan klasifikasi emosi, text processing, dan metode klasifikasi dengan algoritma Naïve Bayes. Informasi yang dibutuhkan diperoleh dari beragam sumber, seperti jurnal, buku, dan literatur lainnya.
3. Identifikasi dan Perumusan Masalah
   Identifikasi dan perumusan masalah dilakukan untuk mengetahui arah dari penelitian akan dilakukan. Berdasarkan dari masalah yang ingin diangkat, lalu juga akan menentukan tujuan dari penelitian.
4. Mengumpulkan Data
   Populasi dalam penelitian ini adalah teks dari tweet yang diposting oleh pengguna di platform media sosial Twitter pada bulan Juli 2023, khususnya pada minggu pertama. Data ini diperoleh melalui proses crawling menggunakan layanan situs web pihak ketiga, Netlytic. Proses crawling dilakukan dengan mencari kata kunci yang terkait dengan nama para bakal calon presiden, yakni Anies Baswedan, Ganjar Pranowo, dan Prabowo Subianto, dengan pembagian 5.000 data tweet terhadap masing-masing calon. Ukuran sampel yang ideal adalah 10% dari populasi [16]. Oleh sebab itu sampel yang akan digunakan adalah 10% dari total keseluruhan data tweet.
5. Data *Preprocessing*
   Dalam data *preprocessing* dilakukan proses persiapan sebelum data dilanjutkan proses klasifikasi dengan tujuan agar meningkatkan kinerja dari proses perhitungan model. Proses preprocessing dilakukan dengan langkah-langkah yang sudah disebutkan pada pada bab sebelumnya.
6. Melabelkan data
   Pada proses ini dilakukan pelabelan terhadap dataset tweet yang didapatkan. Menurut. emosi terbagi menjadi enam kategori, yaitu senang, sedih, marah, takut, terkejut, jijik. Dataset akan diberikan label berdasarkan enam kategori emosi tersebut, ditambah dengan satu label emosi netral, total dari pelabelan data yang akan digunakan adalah tujuh label.
7. Ekstraksi Fitur dengan TF-IDF
   Pada tahapan ini dilakukan proses menghitung bobot setiap kata dalam suatu dokumen dan menentukan kerelevanan dari tiap kata tersebut. Algoritma ini digunakan untuk menganalisa hubungan antara dokumen dan memilih fitur bobot kata yang relevan. Fitur-fitur yang akan diekstraksi dari TF-IDF sendiri antara lain yaitu bobot kata, matriks TF-IDF, dan kata-kata kunci atau kata dasar yang penting dalam dokumen.[17]
8. Pembagian Data Latih dan Data Uji
   Tahap ini dilakukan sebagai pembagian data untuk memulai proses klasifikasi, dimulai dengan melabeli data menjadi 6 label berdasarkan kategori emosi yang disebutkan sebelumnya. Langkah berikutnya melibatkan pembagian data menjadi dua kelompok, yakni data latih dan data uji. Pembagian ini mengikuti prinsip Pareto, yang umumnya dikenal

Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika dan Komputer)
Volume 23 ; Nomor 1 ; Februari 2024 ; Page 37-45
E-ISSN : 2615-3475; P-ISSN : 1978-6603
https://ojs.trigunadharma.ac.id/index.php/jis/index

sebagai pembagian 80/20. Jumlah data latih dibagi sebesar 80%, sementara data uji dibagi sebesar 20%. Dalam tahap evaluasi, model aturan 80/20 memiliki kelebihan representasi yang baik terutama dalam mengolah data yang cukup besar untuk mempelajari pola dari suatu data

9. Klasifikasi Emosi dengan Naïve Bayes
   Pada tahap ini dilakukan proses penginputan data latih dan dilakukan analisis dengan Naïve Bayes Classifier yang mana nantinya akan dilanjutkan dengan proses pengujian menggunakan data uji.
10. Evaluasi
    Hasil dari klasifikasi kemudian dilanjutkan pada tahap evaluasi. Tahap evaluasi menggunakan confusion matrix untuk menghitung nilai akurasi, presisi dan sensitivitas.

3. HASIL DAN PEMBAHASAN

3.1 *Preprocessing* data

Data yang telah didapatkan tidak bisa langsung digunakan, data yang masih berbentuk mentah harus dilakukan preprocessing data terlebih dahulu. Tujuan dari preprocessing data adalah agar data mentah yang didapatkan nantinya lebih akurat ketika digunakan dalam tahap lanjut analisis data. Terdapat beberapa langkah preprocessing yang dilakukan, yaitu tokenisasi, case folding, stopword removal, penghapusan karakter non-alfanumerik, stemming, dan lemmatization. Preprocessing dataset pada penelitian ini menggunakan bahasa pemrograman Python dan library NLTK.

3.2 Pelabelan Data

Klasifikasi adalah teknik yang biasanya juga dikenal dengan algoritma supervised learning. Sistem diberikan training untuk mempelajari berdasarkan data yang telah ada atau telah dilabeli sebelumnya, kemudian pola yang didapatkan tersebut menjadi acuan untuk kumpulan data berikutnya [18]. Proses pelabelan dilakukan secara manual yang diawasi oleh dosen Bahasa Indonesia Fakultas Bahasa dan Seni Universitas Negeri Medan yang ahli dalam bidang psikolinguistik.

Data *tweet* bakal calon Anies Baswedan, terdapat sebanyak 484 *tweet* untuk emosi senang, 381 *tweet* untuk emosi marah, 232 *tweet* untuk emosi netral, 41 *tweet* emosi jijik, 36 *tweet* emosi sedih, 32 *tweet* emosi takut, dan 4 *tweet* untuk emosi terkejut.

Untuk pelabelan emosi terhadap data *tweet* bakal calon Ganjar Pranowo, terdapat sebanyak 668 *tweet* untuk emosi senang, 273 *tweet* untuk emosi netral, 70 *tweet* untuk emosi marah, 40 *tweet* emosi terkejut, 21 *tweet* emosi takut, 18 *tweet* emosi jijik, dan 9 *tweet* untuk emosi sedih.

Adapun pelabelan emosi terhadap data *tweet* bakal calon Prabowo Subianto, terdapat sebanyak 638 *tweet* untuk emosi senang, 468 *tweet* untuk emosi netral, 126 *tweet* untuk emosi marah, 111 *tweet* emosi terkejut, 72 *tweet* emosi sedih, 54 *tweet* emosi takut, dan 11 *tweet* untuk emosi jijik.

3.3 Ekstraksi Fitur dengan TF-IDF

Term Frequency-Inverse Document Frequency (TD-IDF) adalah teknik yang umum digunakan dalam pemrosesan bahasa alami untuk mewakili teks dalam bentuk fitur numerik yang dapat digunakan dalam model machine learning. TF-IDF memberikan bobot pada kata-kata berdasarkan seberapa sering kata tersebut muncul di dokumen dan seberapa penting kata tersebut dalam seluruh kumpulan dokumen.

Tabel 1. Contoh Skenario Ekstraksi Fitur TF-IDF

Term	**TF**			**DF**	**IDF** $\frac{Log\ n}{DF+1}$	**TF-IDF** *(tf x idf)*		
	1	2	1481			1	2	1481
Mesra	$\frac{1}{20} = 0.05$	0	0	1	0.176	0.0088	0	0
Prabowo	$\frac{2}{20} = 0.1$	$\frac{1}{7} = 0.14$	$\frac{1}{6} = 0.16$	3	-0.124	-0.012	0.017	0.019
Indah	0	$\frac{1}{7} = 0.14$	0	1	0.176	0	0.024	0

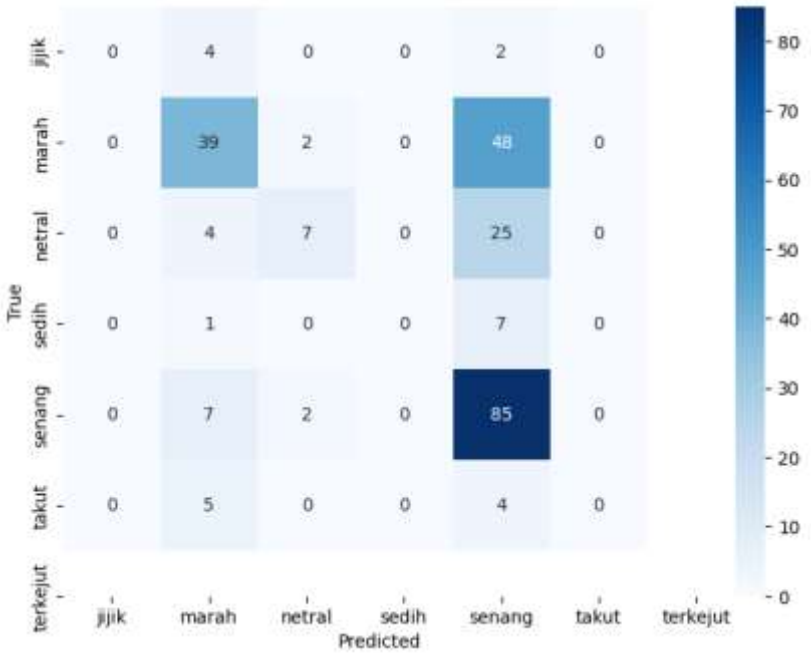
Pada bahasa *Python* untuk melakukan proses TF-IDF dapat menggunakan *TfidfVectorizer* yang berasal dari *library Scikit-Learn.*

3.4 Pengujian dan Evaluasi Model

Proses pengklasifikasian dan pengujian model dalam penelitian menggunakan bahasa pemrograman python dengan menggunakan fungsi MultinomialNB dan proses evaluasi menerapkan confusion matrix untuk menampilkan nilai akurasi pada model. Proses pengujian dilakukan dua kali terhadap tiap calon, dilakukan dengan menggunakan label netral, dan tanpa label netral. Tujuan dari pelabelan tersebut untuk melihat apakah ada pengaruh label netral terhadap data. Hasil akurasi yang ditampilkan pada tiap calon memiliki jumlah yang berbeda. Hasil Klasifikasi tiap calon dapat dilihat pada tabel 1, 2, dan 3.

Tabel 2. Hasil Klasifikasi Data Tweet Anies Baswedan

Label	*Precision*	*Recall*	*F1-score*	*Support*
Jijik	0,00	0,00	0,00	6
Marah	0,69	0,45	0,54	89
Netral	0,58	0,19	0,29	36
Sedih	0,00	0,00	0,00	8
Senang	0,51	0,94	0,66	94
Takut	0,00	0,00	0,00	9
Terkejut	0,00	0,00	0,00	0
Akurasi	0.5578512396694215			



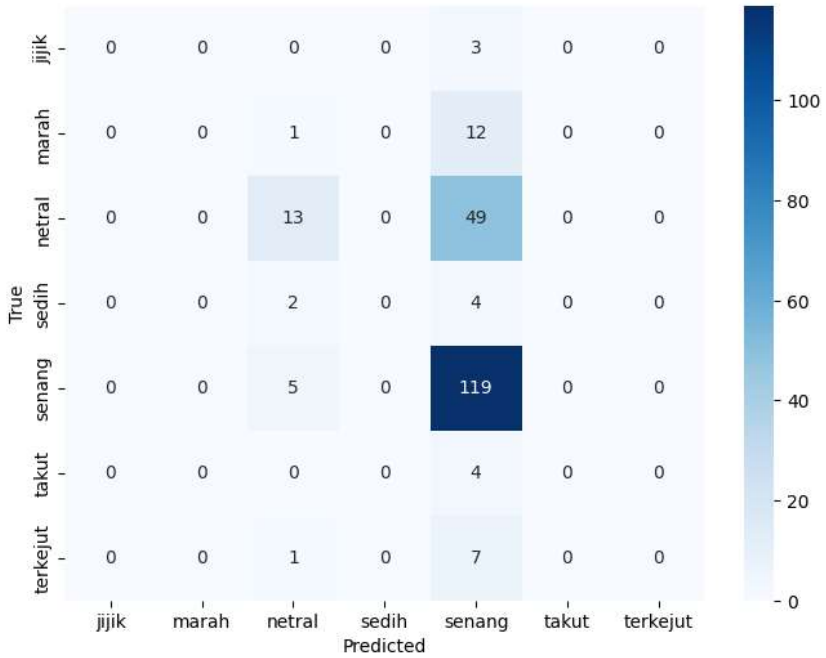
Gambar 2. *Confusion Matrix* Data Tweet Anies Baswedan

Pada gambar 2 *output* pengujian model pada data *tweet* Anies Baswedan, dengan perbandingan data 80:20 dan menggunakan kelas netral. Memiliki akurasi sebesar 55,78% dan kelas marah memiliki nilai presisi paling tinggi sebesar 69%, kelas netral 58%, dan senang 51%. Hasil klasifikasi pada gambar 2 menunjukkan ada 85 data senang yang diprediksi benar senang, 39 data marah yang diprediksi benar marah, dan ada beberapa data label emosi lainnya yang diprediksi salah.

Tabel 3. Hasil Klasifikasi Data Tweet Ganjar Pranowo

Label	*Precision*	*Recall*	*F1-score*	*Support*
Jijik	0,00	0,00	0,00	0

Marah	0,00	0,00	0,00	13
Netral	0,59	0,21	0,31	62
Sedih	0,00	0,00	0,00	6
Senang	0,60	0,96	0,74	124
Takut	0,00	0,00	0,00	4
Terkejut	0,00	0,00	0,00	8
Akurasi	0, 6			



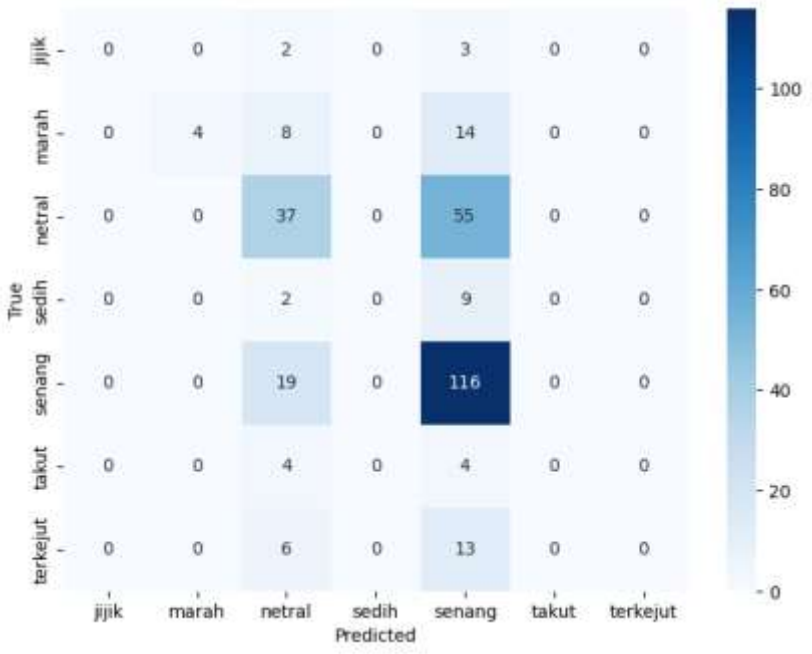
Gambar 3 *Confusion Matrix* Ganjar Pranowo Menggunakan Label Netral

Hasil klasifikasi dari Ganjar Pranowo memiliki nilai akurasi sebesar 60% dengan nilai *precision* label senang paling tinggi, sebanyak 60%. Hasil *confusion matrix* Ganjar Pranowo menunjukkan pada label senang memiliki total 199 data diprediksi benar.

Tabel 4. Hasil Klasifikasi Data Tweet Prabowo Subianto

Label	*Precision*	*Recall*	*F1-score*	*Support*
Jijik	0,00	0,00	0,00	5
Marah	1,00	0,15	0,27	26
Netral	0,47	0,40	0,44	92
Sedih	0,00	0,00	0,00	11
Senang	0,54	0,86	0,66	135
Takut	0,00	0,00	0,00	8
Terkejut	0,00	0,00	0,00	19
Akurasi	0,530			

Hasil klasifikasi pada data *tweet* Prabowo Subianto mengunakan kelas netral menunjukkan label marah memiliki nilai *precision* paling tinggi sebanyak 100%. Label senang memiliki nilai *precision* 54%.



Gambar 4. *Confusion Matrix* Prabowo Subianto

Terdapat 116 data label senang yang diprediksi benar, lalu 37 data label netral yang diprediksi benar. Terdapat juga beberapa data yang mengalami prediksi yang salah.

4. KESIMPULAN

Berdasarkan penelitian dan analisis yang telah dilakukan, dapat disimpulkan bahwa, terdapat buzzer yang melakukan tweet yang sama berulang-ulang terhadap masing-masing bakal calon presiden dengan tujuan mendongkrak citra tiap calon. Hasil dari pra-processing cleaning dan filterasi data terhadap buzzer tiap bakal calon presiden, awalnya terdapat masing-masing 5000 data tweet hasil crawling, namun hasil akhir data tweet masing-masing bakal calon presiden setelah dilakukan proses filterasi yaitu, Anies Baswedan (1211 data tweet), Ganjar Pranowo (1101 data tweet), Prabowo Subianto (1481 data tweet). Hasil akhir tersebut menunjukkan bahwa Ganjar Pranowo memiliki tweet buzzer paling banyak, disusul dengan Anies Baswedan, dan Prabowo Subianto.

Hasil dari pelabelan emosi dari tiap bakal calon presiden menunjukkan nilai yang berbeda satu sama lain, adapun jumlah pelabelan emosi masing-masing calon yaitu : Anies Baswedan (senang : 484, marah : 381, netral : 232, jijik : 41, sedih : 36, takut : 32, terkejut : 4), Ganjar Pranowo (senang : 668, netral : 274, marah : 70, terkejut : 40, takut : 21, jijik : 18, sedih : 9), Prabowo Subianto (senang : 638, netral : 468, marah : 126, terkejut : 111, sedih : 72, takut : 54, jijik : 11).

Berdasarkan penelitian yang dilakukan juga, performa model algoritma Naïve Bayes tanpa menggunakan label netral menunjukkan performa yang lebih baik, dengan nilai akurasi model sebesar 58% pada data Anies Baswedan, 58% pada data Prabowo Subianto, dan 76% pada data Ganjar Pranowo, serta gabungan 69%.

Sedangkan pada performa model menggunakan label netral menunjukkan adanya pengaruh label netral terhadap akurasi dari model, hal ini terjadi karena label netral mendominasi kelas emosi lainnya. Klasifikasi pada skenario pengujian menggunakan label netral menunjukkan akurasi sebesar 55% pada data Anies Baswedan, 60% pada data Ganjar Pranowo, dan 53% pada data Prabowo Subianto, sedangkan untuk gabungan semuanya nilai akurasi sebesar 51%.

Untuk mengatasi ketidakseimbangan data dilakukan *random sampling*. Tahap *random sampling* dilakukan bertujuan untuk menyeimbangkan seluruh data dalam label, sehingga model dapat memberikan hasil yang lebih baik. Hasil klasifikasi random sampling menunjukkan nilai akurasi 44% untuk data Anies menggunakan label netral, dan 44% tanpa label netral. Ganjar Pranowo menunjukkan nilai akurasi 71% menggunakan label netral, dan 71% tanpa label netral. Prabowo Subianto memberikan nilai akurasi model 43% menggunakan label netral, dan 52% tanpa label netral. Terdapat beberapa kelas emosi yang tidak dapat diklasifikasikan dengan fscore di bawah satu persen, karena sedikitnya jumlah data terkait kelas tersebut.

DAFTAR PUSTAKA

[1] A. Nizar, "KLASIFIKASI EMOSI PADA CUITAN DI TWITTER DENGAN PRINCIPAL COMPONENT ANALYSIS DAN SUPPORT VECTOR MACHINE," vol. 10, no. 01, pp. 13–20, 2022.

[2] R. Septian *et al.*, "Klasifikasi Emosi Menggunakan Convolutional Neural Networks Emotion Classification Based on Convolutional Neural Networks," no. October, pp. 53–62, 2020.

[3] R. Feldman and J. Sanger, *The text mining handbook: Advanced approaches in analyzing unstructured data*. 2007.

[4] Karsito and S. Susanti, "Klasifikasi Kelayakan Peserta Pengajuan Kredit Rumah Dengan Algoritma Naïve Bayes Di

Perumahan Azzura Residencia," *J. Teknol. Pelita Bangsa*, vol. 9, pp. 43–48, 2019.
[5] I. Werdiningsih, D. Novitasari, and D. Haq, *PENGELOLAAN DATA MINING DENGAN PEMROGRAMAN MATLAB*. Surabaya: Airlangga University Press, 2022.
[6] Kusrini and E. Luthfi, *Algoritma Data Mining*. Yogyakarta: ANDI, 2019.
[7] A. Ifon Purnama, A. Aziz, A. Sartika Wiguna, and K. Kunci, "Penerapan Data Mining Untuk Mengklasifikasi Penerima Bantuan PKH Desa Wae Jare Menggunakan Metode Naïve Bayes," *Kurawal J. Teknol. Inf. dan Ind.*, vol. 3, pp. 1–8, 2020.
[8] A.- Arini, L. K. Wardhani, and D.- Octaviano, "Perbandingan Seleksi Fitur Term Frequency & Tri-Gram Character Menggunakan Algoritma Naïve Bayes Classifier (Nbc) Pada Tweet Hashtag #2019gantipresiden," *Kilat*, vol. 9, no. 1, pp. 103–114, 2020, doi: 10.33322/kilat.v9i1.878.
[9] S. D. Harijiatno, "Analisis Sentimen pada Twitter Menggunakan Multinominal Naive Bayes," 2019.
[10] K. Aulia and L. Amelia, "Analisis Sentimen Twitter Pada Isu Mental Health Dengan Algoritma Klasifikasi Naive Bayes," *Siliwangi J. (Seri Sains Teknol.*, vol. 6, no. 2, pp. 60–65, 2020.
[11] Muhammad Romzi and B. Kurniawan, "Pembelajaran Pemrograman Python Dengan Pendekatan Logika Algoritma," *JTIM J. Tek. Inform. Mahakarya*, vol. 03, no. 2, pp. 37–44, 2020.
[12] R. M. R. Clinton and S. Sengkey, "Purwarupa Sistem Daftar Pelanggaran Lalulintas Berbasis Mini-Komputer Raspberry Pi," *J. Tek. Elektro dan Komput. Vol.8*, vol. 8, no. 3, pp. 181–192, 2019.
[13] A. C. Herlingga, I. P. E. Prismana, D. R. Prehanto, and D. A. Dermawan, "Algoritma Stemming Nazief & Adriani dengan Metode Cosine Similarity untuk Chatbot Telegram Terintegrasi dengan E-layanan," *J. Informatics Comput. Sci.*, vol. 2, no. 01, pp. 19–26, 2020, doi: 10.26740/jinacs.v2n01.p19-26.
[14] R. D. Himawan and E. Eliyani, "Perbandingan Akurasi Analisis Sentimen Tweet terhadap Pemerintah Provinsi DKI Jakarta di Masa Pandemi," *J. Edukasi dan Penelit. Inform.*, vol. 7, no. 1, p. 58, 2021, doi: 10.26418/jp.v7i1.41728.
[15] Sutopo, "PENENTUAN JUMLAH SAMPEL DALAM PENELITIAN," vol. 21, no. 1, pp. 1–9, 2020, [Online]. Available: http://journal.um-surabaya.ac.id/index.php/JKM/article/view/2203.
[16] C. R. Mirsandi, "Implementasi program keluarga harapan (pkh) dalam memberikan perlindungan sosial pada masyarakat (studi dikecamatan setia kabupaten aceh barat daya)," *J. Chem. Inf. Model.*, pp. 1–103, 2019.
[17] B. Brahimi, M. Touahria, and A. Tari, "Improving sentiment analysis in Arabic: A combined approach," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 33, no. 10, pp. 1242–1250, 2021, doi: 10.1016/j.jksuci.2019.07.011.
[18] H. Abijono, P. Santoso, and N. L. Anggreini, "Algoritma Supervised Learning Dan Unsupervised Learning Dalam Pengolahan Data," *J. Teknol. Terap. G-Tech*, vol. 4, no. 2, pp. 315–318, 2021, doi: 10.33379/gtech.v4i2.635.