Jurnal Teknik Informatika (JUTIF)
Vol. 5, No. 1, February 2024, pp. 321-329

DOI: https://doi.org/10.52436/1.jutif.2024.5.1.1817
p-ISSN: 2723-3863
e-ISSN: 2723-3871

COMPARISON OF RANDOM FOREST, SUPPORT VECTOR MACHINE AND NAIVE BAYES ALGORITHMS TO ANALYZE SENTIMENT TOWARDS MENTAL HEALTH STIGMA

Putri Elisa[*1], Auliya Rahman Isnain[2]

[1,2]Information Systems, Faculty of Engineering and Computer Science, Universitas Teknokrat Indonesia, Indonesia
Email: [1]putri_elisa@teknokrat.ac.id, [2]auliyarahman@teknokrat.ac.id

(Article received: February 01, 2024; Revision: February 17, 2024; published: February 24, 2024)

Abstract

Advances in technology, especially the internet, have significantly changed the way people communicate, including social media. Social media facilitates more effective and efficient online communication. Twitter has 18.45 million users in Indonesia by 2022. Discussion of mental health stigma on twitter, increased 17% in 2021 compared to the previous year. Lifestyle transformation, social pressures, and technological advancements have created new challenges in maintaining individual mental health. Discussions of mental health issues have become pros and cons on twitter. The tendency of twitter users in posting content can be known by means of sentiment analysis. Therefore, sentiment analysis can be used to classify comments and tweets related to mental health stigma into negative, positive and neutral. So, it is expected to provide a number of significant benefits in the aspect of managing mental health issues. The methods used to analyze sentiment towards mental health stigma are Random Forest, Support Vector Machine (SVM) and Naïve Bayes algorithms. Based on the research that has been done, it produces 3,095 data for the period 2020-2023. After preprocessing and labeling the data, 1,635 data (negative class), 633 data (positive class) and 208 data (neutral class) were obtained. The SVM model test results show an accuracy of 86.11%, the Random Forest model shows an accuracy of 82.55%, while the Naive Bayes model shows an accuracy of 78.19%. Therefore, it can be concluded that SVM has the best performance in classifying tweets containing mental health stigma.

Keywords: *Mental Health, Naive Bayes, Random Forest, Sentiment Analysis, Support Vector Machine.*

PERBANDINGAN ALGORITMA RANDOM FOREST, SUPPORT VECTOR MACHINE DAN NAIVE BAYES UNTUK MENGANALISIS SENTIMEN TERHADAP STIGMA KESEHATAN MENTAL

Abstrak

Kemajuan teknologi khususnya internet, telah merubah cara masyarakat berkomunikasi secara signifikan termasuk media sosial. Sosial media memfasilitasi komunikasi online yang lebih efektif dan efisien. Twitter memiliki 18,45 juta pengguna di Indonesia pada tahun 2022. Pembahasan stigma kesehatan mental pada twitter, meningkat 17% pada tahun 2021 dibandingkan tahun sebelumnya. Transformasi gaya hidup, tekanan sosial, dan kemajuan teknologi telah menciptakan tantangan baru dalam menjaga kesehatan mental individu. Pembahasan isu kesehatan mental menjadi pro dan kontra di twitter. Kecenderungan pengguna twitter dalam memposting konten dapat diketahui dengan cara analisis sentimen. Oleh karena itu analisis sentimen dapat digunakan untuk mengklasifikasikan komentar dan tweet terkait stigma kesehatan mental menjadi negatif, positif dan netral. Sehingga diharapkan dapat memberikan sejumlah manfaat yang signifikan dalam aspek pengelolaan isu kesehatan mental. Metode yang digunakan untuk menganalisis sentimen terhadap stigma kesehatan mental adalah algoritma Random Forest, Support Vector Machine (SVM) dan Naïve Bayes. Berdasarkan penelitian yang telah dilakukan menghasilkan 3.095 data periode 2020-2023. Setelah melalui praproses dan pelabelan data didapatkan 1.635 data (kelas negatif), 633 data (kelas positif) dan 208 data (kelas netral). 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.

Kata kunci: *Analisis Sentimen, Kesehatan Mental, Naive Bayes, Random Forest, Support Vector Machine.*

1. PENDAHULUAN

Kemajuan teknologi, khususnya internet, telah merubah cara masyarakat berkomunikasi secara signifikan. Berbagai platform, termasuk media sosial, telah diperkenalkan untuk memfasilitasi komunikasi online yang lebih efektif dan efisien [1]. Menariknya, Indonesia mencatatkan 191 juta pengguna aktif media sosial pada bulan Januari 2022 dan mengalami peningkatan signifikan sebesar 12,35% dari tahun sebelumnya. Twitter, sebagai salah satu platform terkemuka, memiliki 18,45 juta pengguna di Indonesia pada tahun 2022 [2].

Sebagai platform komunikasi yang mudah diakses kapan pun, media sosial twitter akhirnya memiliki dampak yang signifikan pada pertukaran informasi di antara penggunanya. Salah satu fokus pembicaraan yang tidak terhindarkan di media sosial twitter adalah isu kesehatan [3]. Selah satu bentuk mewujudkan kesehatan secara menyeluruh adalah dengan menjaga kesehatan mental [4].

Isu-isu seputar kesehatan mental menjadi perbincangan yang menarik di kalangan pengguna twitter, terutama dari generasi milenial dan Gen-Z, sebagai bentuk kepedulian mereka terhadap kesehatan mental [5]. Berdasarkan data yang dikumpulkan dari platform Twitter pada tahun 2021, terlihat adanya peningkatan sebesar 17% dalam tingkat pembicaraan dan diskusi mengenai aspek kesehatan mental dari tahun 2018-2021 [6].

Kesehatan mental merujuk pada kondisi kejiwaan atau psikologis seseorang yang mencerminkan kemampuannya untuk beradaptasi dan mengatasi tantangan baik yang berasal dari dirinya sendiri (internal) maupun dari lingkungan eksternalnya [7]. Berdasarkan data WHO tahun 2016, diperkirakan sekitar 35 juta individu mengalami depresi, 60 juta orang mengidap gangguan bipolar, 21 juta orang menghadapi skizofrenia, dan 47,5 juta orang menderita demensia [8].

Pada era modern seperti saat ini, kesehatan mental telah menjadi sorotan utama. Transformasi gaya hidup, tekanan sosial, dan kemajuan teknologi telah menciptakan tantangan baru dalam menjaga kesehatan mental individu. Perubahan dalam pola hidup yang terkait dengan urbanisasi, globalisasi, dan perkembangan teknologi membawa beban baru yang menimbulkan tantangan psikologis yang sebelumnya belum pernah dihadapi [9]. Maka dibutuhkan pemahaman mendalam tentang stigma terkait kesehatan mental agar dapat mendukung perancangan strategi pencegahan gangguan mental dan intervensi yang efektif, serta memastikan bahwa setiap individu menerima bantuan penanganan kesehatan mental yang tepat.

Pembahasan isu kesehatan mental menjadi pro dan kontra di twitter. Hal ini menyebabkan munculnya fenomena perdebatan di twitter yang sebenarnya menunjukkan perhatian mengenai isu kesehatan mental tersebut [10]. Kecenderungan pengguna twitter dalam memposting konten dapat diketahui dengan cara analisis sentimen [11].

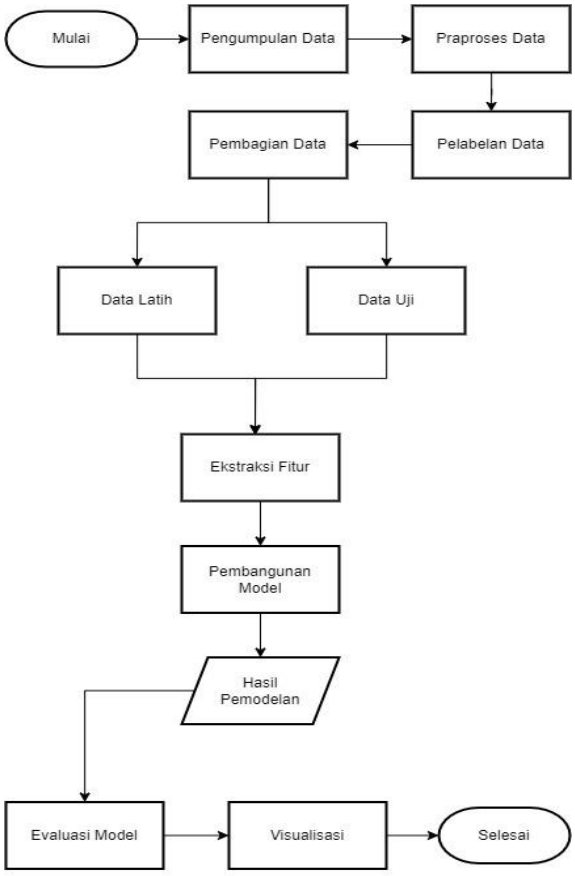
Analisis sentimen ialah proses mengekstraksi, mengolah dan memahami data berupa teks yang tidak terstruktur secara otomatis guna mengambil informasi sentimen yang terdapat pada sebuah kalimat pendapat atau opini [12]. Analisis sentimen dapat diterapkan pada opini semua bidang seperti ekonomi, politik, sosial dan hukum [13]. Oleh karena itu analisis sentimen dapat digunakan untuk mengklasifikasikan komentar dan tweet terkait stigma kesehatan mental menjadi negatif, positif dan netral [14]. Sehingga diharapkan dapat memberikan sejumlah manfaat yang signifikan dalam aspek pengelolaan isu kesehatan mental.

Beberapa penelitian terkait berdasarkan permasalahan dan solusi yang sejenis telah dilakukan, salah satunya penelitian yang menganalisis sentimen komentar netizen twitter terhadap kesehatan mental masyarakat indonesia menggunakan metode naïve bayes. Hasil analisis sentimen mendapat akurasi naïve bayes didapatkan sebesar 79% [15]. Penelitian lain membandingkan algoritma Naive Bayes untuk klasifikasi sentimen masyarakat tentang depresi pada youtube. Hasil analisis sentimen algoritma NBC mendapat akurasi sebesar 84.11% [16]. Selain itu terdapat penelitian lainnya yang membandingkan data mining dalam kasus *mental health* pada sosial media twitter menggunakan metode Naive Bayes. Hasil penelitian menunjukan bahwa algoritma Naïve Bayes memiliki nilai macro average untuk precision, recall, dan f1-score dengan nilai 63% hingga 74% [14]. Selanjutnya penelitian lain melakukan analisis sentimen kemungkinan depresi dan kecemasan pada twitter menggunakan Support Vector Machine. Berdasarkan analisis hasil yang telah dilakukan SVM mendapat akurasi sebesar 82.5% [17]. Berbeda dengan penelitian sebelumnya, terdapat penelitian yang merancang dan membangun *text mining* untuk deteksi kecemasan menggunakan *machine learning* berdasarkan data media sosial selama pandemi COVID-19. Hasil dari penelitian ini mendapat akurasi dari metode Random Forest dan XGBOOST adalah 83% dan 73% [18].

Berdasarkan permasalahan sebelumnya serta beberapa penelitian terkait yang telah disebutkan, penelitian ini bertujuan untuk membandingkan algoritma Random Forest, Support Vector Machine (SVM) dan Naïve Bayes untuk menganalisis sentimen terhadap stigma kesehatan mental. Terdapat beberapa kontribusi yang harapkan dari penelitian ini seperti penelitian ini berbeda dari penelitian sebelumnya yang banyak menggunakan algoritma Naive Bayes. Penelitian ini mengisi celah tersebut dengan membandingkan kinerja SVM, Random Forest dan Naive Bayes dalam menganalisis sentimen stigma kesehatan mental. Selain itu penelitian ini diharapkan dapat membantu mengidentifikasi pola sentimen sehingga dapat digunakan dalam pengelolaan kesehatan mental di masyarakat.

2. METODE PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini didesain untuk memberikan pemahaman mendalam terhadap sentimen terhadap stigma kesehatan yang di ambil pada media sosial twitter dengan membandingkan algoritma Random Forest, SVM dan Naive Bayes. Pada Gambar 1, terdapat urutan langkah-langkah yaitu pengumpulan data, eksplorasi data, praproses data, pelabelan data, pembagian data, ekstraksi fitur, pembangunan model, evaluasi model dan visualisasi hasil prediksi. Metode penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.1. Pengumpulan Data

Pengumpulan data dilakukan menggunakan metode *crawling*. Metode *crawling* adalah suatu teknik pengambilan data dari suatu website secara otomatis dengan menggunakan program komputer. Metode ini memungkinkan pengambilan data dari berbagai sumber secara efisien dan terstruktur, sehingga dapat digunakan untuk berbagai keperluan seperti penelitian, analisis, dan pengembangan aplikasi [19]. Pada penelitian ini data dikumpulkan dari sosial media twitter dengan menggunakan keyword tertentu yang membahas stigma kesehatan mental. Data yang dikumpulkan dalam rentan tahun 2020 hingga 2023. Analisis sentimen penelitian ini menggunakan variabel *full text* yang terdapat pada data stigma kesehatan mental.

2.2. Praproses Data

Data yang digunakan dalam penelitian ini adalah tweet dari platform twitter dalam bentuk teks yang bersifat tidak terstruktur, sehingga dibutuhkan langkah praproses. Teknik praproses data yang digunakan secara berurutan adalah *text cleaning* untuk menghilangkan karakter yang tidak diperlukan, *case folding* digunakan untuk merubah ukuran huruf kapital menjadi huruf kecil, *tokenizing* digunakan untuk memecah kalimat menjadi kata, *normalization* digunakan untuk merubah kata yang tidak baku menjadi kata baku, *stemming* bertujuan mengubah kata menjadi kata dasar dan menghapus imbuhan, sedangkan *filtering* digunakan untuk menghapus kata yang tidak memiliki makna khusus pada kalimat [20].

2.3. Pelabelan Data

Setelah dataset telah terpenuhi, langkah selanjutnya melakukan pelabelan pada setiap tweet tersebut. Pelabelan merupakan proses klasifikasi pada data untuk menentukan apakah setiap kalimat dalam dataset tersebut memiliki makna yang positif, negatif atau netral [9]. Penentuan label dilakukan pada data teks yang memiliki kata pada kamus lexicon [21]. Adapun persamaan pelabelan positif, negatif dan netral [21], dapat dilihat pada persamaan 1.

$$S_{sentiment} = \begin{Bmatrix} positive\ if\ S_{positive} > S_{negative} \\ neutral\ if\ S_{positive} = S_{negative} \\ negative\ if\ S_{positive} < S_{negative} \end{Bmatrix} \quad (1)$$

Dimana:
$S_{sentiment}$: Kalimat sentimen pada dataset
$S_{positive}$: Kalimat sentimen positif pada dataset
$S_{negative}$: Kalimat sentimen negatif pada dataset

2.4. Pembagian Data

Selanjutnya adalah tahapan pembagian data yang merupakan tahapan membagi data set menjadi data latih dan data uji. Pada penelitian ini menggunakan 80% sebagai data latih dan 20% sebagai data uji. Data latih digunakan untuk membangun model berdasarkan opini dari setiap kategori, yaitu positif, negatif dan netral. Sedangkan data uji digunakan untuk menguji model yang telah dibangun agar mendapat model terbaik [22].

2.5. Ekstraksi Fitur

Ekstraksi fitur dilakukan karena *machine learning* tidak dapat memahami karakter dan kata-kata yang ada pada dataset. Sehingga ketika berhadapan dengan data teks perlu direpresentasikan ke dalam bentuk angka agar dapat dipahami. Metode ekstraksi fitur yang digunakan dalam penelitian ini adalah Count-Vectorizer. Count-Vectorizer adalah sebuah ekstraksi fitur yang mengolah dokumen atau teks kemudian direpresentasikan dalam bentuk vektor [23].

2.6. Pembangunan Model

Pada tahap pembangunan pemodelan dilakukan proses klasifikasi data menggunakan tiga algoritma, yaitu Random Forest, Support Vector Machine dan Naive Bayes. Pembangunan model ini menggunakan bahasa pemrograman python dengan library sklearn. Berikut merupakan penjelasan lebih detail dari ketiga algoritma tersebut:

2.6.1. Random Forest

Random Forest merupakan metode yang mampu meningkatkan hasil akurasi dalam membangkitkan atribut untuk setiap node yang dilakukan secara acak [24]. Pemilihan kelas pada Random Forest yang paling populer [25], dapat menggunakan persamaan 2.

$$f(x) = Average\big(f_1(x), f_2(x), \dots, f_n(x)\big) \quad (2)$$

Dimana:
$f(x)$: Hasil prediksi
$f_{1\text{-}n}(x)$: Hasil prediksi dari setiap pohon keputusan ke-n
(x) : Inputan data

2.6.2. Support Vector Machine (SVM)

SVM adalah algoritma klasifikasi berbasis diskriminatif yang bermaksud untuk mencari batasan pemisahan terbaik yang disebut *hyperplane* yang akan memisahkan kelas [26]. *Hyperplane* yang optimal pada linear kernel [17], direpresentasikan pada persamaan 3 dan 4.

$$(w.x_i + b) \leq 1, y_i = -1 \quad (3)$$

$$(w.x_i + b) \geq 1, y_i = 1 \quad (4)$$

Dimana:
x_i : Data ke-i
$w.x_i$: Nilai bobot untuk kelas data ke-i
b : Nilai bias
y_i : Kelas data ke-i

2.6.3. Naïve Bayes

Metode ini memiliki potensi yang baik untuk mengklasifikasi dalam hal akurasi dan perhitungan data. Naïve Bayes Classifer menghitung probabilitas jika kategori keputusan adalah benar karena merupakan vektor informasi objek, yang mengamsusikan bahwa atribut objek ini adalah independent [2]. Adapun teorema bayes [27], dapat dilihat dari persamaan 5.

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (5)$$

Dimana:
X : Data dengan kelas yang belum diketahui
H : Hipotesis data merupakan suatu kelas Spesifik
$P(H|X)$: Probabilitas hipotesis H berdasarkan kondisi X (posteriori probabilitas).
$P(H)$: Probabilitas hipotesis H (prior probabilitas).
$P(X|H)$: Probabilitas X berdasarkan kondisi pada hipotesis H.
$P(X)$: Probabilitas X

2.7. Evaluasi Model

Metode evaluasi yang digunakan pada penelitian ini adalah *confusion matrix*. *Confusion matrix* merupakan sebuah teknik yang mudah dan efektif dalam mengukur kinerja sistem klasifikasi [28]. Pada penelitian ini, terdapat tiga kelas sentimen yaitu positif, negatif dan netral. *Confusioan matrix* digunakan untuk menghitung nilai *accuracy*, *precision*, *recall* dan *f1-score*.

3. HASIL DAN PEMBAHASAN

3.1. Pengumpulan Data

Pengumpulan data menggunakan metode *crawling* tersebut diterapkan pada sosial media twitter. Adapun untuk keyword yang digunakan dalam pengumpulan data diantaranya "kesehatan mental", "kesehatan mental gen z", "mental gen z" dan "mental health". Jumlah data tweet yang berhasil dikumpulkan adalah 3.095 *records.* Data yang dikumpulkan dalam rentan tahun 2020 hingga 2023. Kemudian data disimpan dalam format csv sebelum digunakan dalam analisis sentimen. Crawling data tersebut menggunakan tools Google Colab dengan bahasa pemograman Python dan library pandas sebagai dasar perintah untuk penambangan datanya. Berikut merupakan dataset yang digunakan dalam pemodelan, dapat dilihat pada tabel 1.

Tabel 1. Sampel Dataset Awal

full_text
@Kudapanangin @tanyarlfes Gen z msh jd zigot, blm tau apa itu mental health.
@ShiruShiru23 @raspeuberry @convomfs jgn takut Kaaaak ku yakin ortu2 yg Millennial &amp Gen Z ga akan kyk mereka &amp ank2 kita nanti bakal lbh bahagia. kit2 tuh generasi yg sudah tercerahkan ngerti soal inner child, mental health, ilmu parenting dll. Anggep aja ortu2 kita dan yg tua2 dulu mah masi hdp zaman Jahiliyah.
@manusiaaia @dani_dosed @darkjok_reborn @_Hrst @ozarangkuti Aowkaowkoakwo tipikal pengikutnya ozarangkuu, bawa bawa mental health gen z dsb dsb yg diulang ulang terus macem burung beo.
@recehtapisayng Masalahnya kalo gen z nikah, ada masalah dikit langsung ngmongin mental health. Gua butuh bahagia.
@ozarangkuti Gen z lebih peduli habbit dan mental health nya terjaga daripada perang ras.

3.2. Praproses Data

3.2.1. *Text Cleaning*

Pada tahapan cleansing, karakter-karakter yang kurang penting akan dihapus atau dihilangkan. Karakter-karakter tersebut meliputi tanda baca,

simbol, emoticon, URL, hastag, karakter, dan angka. Sampel hasil *text cleaning* dapat dilihat pada tabel 2.

Tabel 2. Sampel Hasil *Text Cleaning*

text cleaning
Gen z msh jd zigot blm tau apa itu mental health
jgn takut Kaaaak ku yakin ortu yg Millennial amp Gen Z ga akan kyk mereka amp ank kita nanti bakal lbh bahagia kit tuh generasi yg sudah tercerahkan ngerti soal inner child mental health ilmu parenting dll Anggep aja ortu kita dan yg tua dulu mah masi hdp zaman Jahiliyah

3.2.2. *Case Folding*

Case folding dengan tujuan untuk mengubah teks atau kalimat yang mengandung huruf kapital menjadi *lowercase*. Sampel hasil *case folding* dapat dilihat pada tabel 3.

Tabel 3. Hasil *Case Folding*

text case folding
gen z msh jd zigot blm tau apa itu mental health
jgn takut kaaaak ku yakin ortu yg millennial amp gen z ga akan kyk mereka amp ank kita nanti bakal lbh bahagia kit tuh generasi yg sudah tercerahkan ngerti soal inner child mental health ilmu parenting dll anggep aja ortu kita dan yg tua dulu mah masi hdp zaman jahiliyah

3.2.3. *Tokenizing*

Tokenizing pada tahap ini proses memecah kalimat atau teks menjadi kata-kata individual. Dengan melakukan *tokenizing* lebih mudah menghitung frekuensi kemunculan setiap kata dalam teks tersebut. Sampel hasil *tokenizing* dapat dilihat pada tabel 4.

Tabel 4. Sampel Hasil *Tokenizing*

text tokenizing
gen, z, msh, jd, zigot, blm, tau, apa, itu, mental, health
jgn, takut, kaaaak, ku, yakin, ortu, yg, millennial, amp, gen, z, ga, akan, kyk, mereka, amp, ank, kita, nanti, bakal, lbh, bahagia, kit, tuh, generasi, yg, sudah, tercerahkan, ngerti, soal, inner, child, mental, health, ilmu, parenting, dll, anggep, aja, ortu, kita, dan, yg, tua, dulu, mah, masi, hdp, zaman, jahiliyah

3.2.4. *Normalization*

Normalization adalah pengubahan kata-kata yang tidak baku ke dalam bentuk kata baku. Kata-kata tidak baku dapat berupa singkatan atau kata-kata slang. Sampel hasil *normalization* dapat dilihat pada tabel 5.

Tabel 5. Sampel Hasil *Normalization*

text *normalization*
gen, saja, masih, jadi, zigot, belum, tau, apa, itu, mental, health
jangan, takut, kaaaak, ku, yakin, orang tua, yang, millennial, amp, gen, saja, enggak, akan, kayak, mereka, amp, anak, kita, nanti, bakal, lebih, bahagia, kit, tuh, generasi, yang, sudah, tercerahkan, mengerti, soal, inner, child, mental, health, ilmu, parenting, dll, anggep, saja, orang tua, kita, dan, yang, tua, dulu, mah, masih, hidup, zaman, jahiliyah

3.2.5. *Stemming*

Stemming digunakan untuk menghilangkan imbuhan-imbuhan yang ada dalam sebuah kata sehingga didapatkan kata dasar. Sampel hasil *stemming* dapat dilihat pada tabel 6.

Tabel 6. Sampel Hasil *Stemming*

text stemming
gen, saja, masih, jadi, zigot, belum, tau, apa, itu, mental, health
jangan, takut, kaaaak, ku, yakin, orang tua, yang, millennial, amp, gen, saja, enggak, akan, kayak, mereka, amp, anak, kita, nanti, bakal, lebih, bahagia, kit, tuh, generasi, yang, sudah, cerah, erti, soal, inner, child, mental, health, ilmu, parenting, dll, anggep, saja, orang tua, kita, dan, yang, tua, dulu, mah, masih, hidup, zaman, jahiliyah

3.2.6. *Filtering*

Pada tahap *filtering* akan ditentukan apakah sebuah kata akan digunakan atau dibuang. Daftar stoplist akan dibuat sebelum melakukan proses stopword removal, jika kata-kata terdapat dalam daftar stoplist, maka kata tersebut akan dihapus, sehingga kata-kata yang tersisa akan dianggap kata yang mencirikan isi suatu dokumen. Sampel hasil *filtering* dapat dilihat pada tabel 7.

Tabel 7. Sampel Hasil *Filtering*

text filtering
gen, zigot, mental, health
takut, kaaaak, ku, orang tua, millennial, gen, kayak, anak, bahagia, kit, generasi, cerah, erti, inner, child, mental, health, ilmu, parenting, dll, anggep, orang tua, tua, mah, hidup, zaman, jahiliyah

3.3. Pelabelan Data

Kumpulan data yang telah melalui tahapan praproses, maka selanjutnya adalah tahapan pelabelan data. Pada penelitian ini dilakukan perhitungan *polarity score* pada jumlah yang terdeteksi berdasarkan kamus lexicon sehingga didapatkan sentimen untuk label negatif, positif dan netral. Kemudian tahap selanjutnya yaitu melihat hasil pelabelan sentimen berdasarkan *polarity score* yang diperoleh. Pelabelan sentimen dilakukan kedalam tiga kelas sentimen yaitu sentimen negatif, netral, dan positif. Adapun sampel hasil pelabelan data stigma kesehatan mental pada penelitian ini dapat dilihat pada tabel 8.

Tabel 8. Sampel Hasil Pelabelan

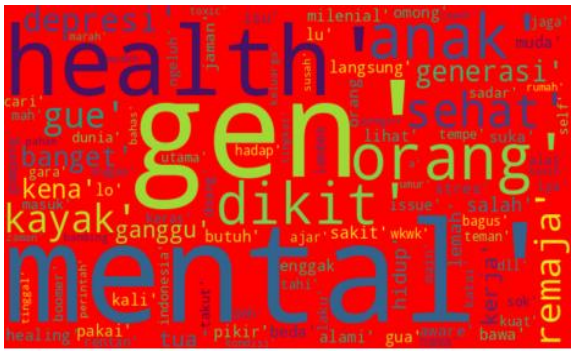
text filtering	polarity score	sentimen
gen, zigot, mental, health	0	Neutral
takut, kaaaak, ku, orang tua, millennial, gen, kayak, anak, bahagia, kit, generasi, cerah, erti, inner, child, mental, health, ilmu, parenting, dll, anggep, orang tua, tua, mah, hidup, zaman, jahiliyah	-11	Negative
work, call, for, all, gen, employees, haloo, sedia, isi, kuesioner, mental, health, gen, kerja, kriteria, gen, didik, min, smasmk, kerja, min, reply, butuh, thankyou	3	Positive

3.3.1. *Wordcloud* kelas sentimen

Wordcloud kelas sentimen digunakan untuk memvisualisasikan frekuensi kata-kata dalam teks

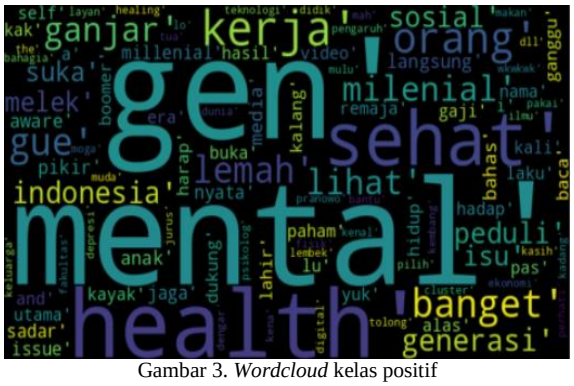
dengan memperbesar kata-kata yang paling sering muncul berdasarkan sentimen kelas negatif, positif dan netral.

Hasil *wordcloud* kelas negatif menampilkan kata yang sering muncul adalah kata “depresi”, “health”, “gen”, “mental” dan “anak”. Adapun hasil dari wordcloud kelas negatif dapat dilihat pada gambar 2.



Gambar 2. *Wordcloud* kelas negatif

Hasil *wordcloud* kelas positif menampilkan kata yang sering muncul adalah kata “kerja”, “health”, “gen”, “mental” dan “sehat”. Adapun hasil dari wordcloud kelas positif dapat dilihat pada gambar 3.



Gambar 3. *Wordcloud* kelas positif

Hasil *wordcloud* kelas positif menampilkan kata yang sering muncul adalah kata “sehat”, “health”, “gen”, “mental” dan “generasi”. Adapun hasil dari wordcloud kelas positif dapat dilihat pada gambar 4.

Gambar 4. *Wordcloud* kelas netral.

3.4. Pembagian Data

Kemudian dataset hasil dari tahapan pelabelan dibagi menjadi data latih dan data uji. Sebelum dilakukan pembagian data, dilakukan proses label encode untuk variabel target yaitu Indonesia Sentiment. Selanjutnya proporsi data latih sebanyak 80% dan data uji sebanyak 20% dari total dataset dengan nilai random state sebesar 42. Sehingga total dari masing-masing dataset menjadi 1.980 data sebagai data latih dan 495 data sebagai data uji.

3.5. Ekstraksi Fitur

Count vectorizer melakukan pembobotan kata pada dengan menghitung frekuensi kemunculan kata dari setiap teks. Kemudian direpresentasikan ke dalam bentuk vektor. Kemudian distribusi kelas hasil ektraksi fitur terdiri dari kelas negatif berjumlah 1.635 data, kelas positif berjumlah 633 data dan kelas netral berjumlah 208 data.

3.6. Pembangunan Model

Pelatihan model Random Forest, SVM dan Naïve Bayes menggunakan data pelatihan hasil ekstraksi fitur. Setelah melatih model, selanjutnya model tersebut untuk membuat prediksi pada data uji.

3.7. Evaluasi Model

3.7.1. Random Forest

Evaluasi model Random Forest mendapatkan presentase precision tertinggi yaitu kelas negatif (88%). Presentase recall tertinggi yaitu kelas negatif (92%). Presentase F1-Score tertinggi yaitu kelas negatif (90%). Presentase akurasi keseluruhan model Random Forest adalah (82,55%). Adapun hasil evaluasi model Random Forest dapat dilihat pada gambar 5.

```
Overall Accuracy Random Forest Model: 0.8255250403877221

Classification Report Random Forest Model:
              precision    recall  f1-score   support

    Negative       0.88      0.92      0.90       420
     Neutral       0.45      0.58      0.51        50
    Positive       0.84      0.64      0.73       149

    accuracy                           0.83       619
   macro avg       0.72      0.71      0.71       619
weighted avg       0.83      0.83      0.82       619
```

Gambar 5. Hasil evaluasi model Random Forest

3.7.2. Support Vector Machine (SVM)

Evaluasi model SVM mendapatkan presentase precision tertinggi yaitu kelas negatif (93%). Presentase recall tertinggi yaitu kelas negatif (92%). Presentase F1-Score tertinggi yaitu kelas negatif (92%). Presentase akurasi keseluruhan model SVM adalah (86,11%). Adapun hasil evaluasi model SVM dapat dilihat pada gambar 6.

```
Overall Accuracy SVM Model: 0.8610662358642972

Classification Report SVM Model:
              precision    recall  f1-score   support

    Negative       0.93      0.92      0.92       420
     Neutral       0.58      0.64      0.61        50
    Positive       0.78      0.78      0.78       149

    accuracy                           0.86       619
   macro avg       0.76      0.78      0.77       619
weighted avg       0.86      0.86      0.86       619
```

Gambar 6. Hasil evaluasi model SVM

3.7.3. Naïve Bayes

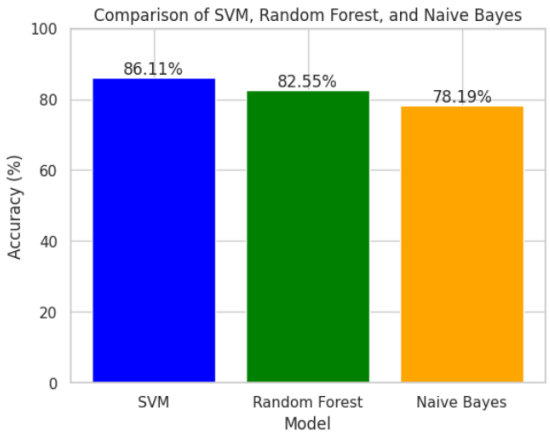
Evaluasi model Naïve Bayes mendapatkan presentase precision tertinggi yaitu kelas negatif (79%). Presentase recall tertinggi yaitu kelas negatif (96%). Presentase F1-Score tertinggi yaitu kelas negatif (87%). Presentase akurasi keseluruhan model Naïve Bayes adalah (78,19%). Adapun hasil evaluasi model Naïve Bayes dapat dilihat pada Gambar 7.

```
Overall Accuracy Naive Bayes Model: 0.7819063004846527

Classification Report Naive Bayes Model:
              precision    recall  f1-score   support

    Negative       0.79      0.96      0.87       420
     Neutral       0.67      0.08      0.14        50
    Positive       0.77      0.51      0.61       149

    accuracy                           0.78       619
   macro avg       0.74      0.52      0.54       619
weighted avg       0.77      0.78      0.75       619
```

Gambar 7. Hasil evaluasi model Naive Bayes

3.8. Visualisasi Perbandingan Model

Visualisasi perbandingan model pada penelitian ini digunakan untuk menampilkan perbandingan akurasi dari model Random Forest, SVM dan Naïve Bayes. Model SVM mendapatkan akurasi sebesar (86,11%). Kemudian model Random Forest mendapatkan akurasi sebesar (82,55%). Sedangkan model Naïve Bayes mendapatkan akurasi sebesar (78,19%). Sehingga yang mendapatkan akurasi tertinggi dari ketiga model tersebut adalah model SVM. Adapun perbandingan akurasi model dapat dilihat pada gambar 8.



Gambar 8. Visualisasi perbandingan akurasi model

4. DISKUSI

Sebelumnya terdapat beberapa penelitian yang melakukan analisis terhadap isu kesehatan mental. Salah satunya adalah penelitian yang menganalisis sentimen kemungkinan depresi dan kecemasan pada twitter menggunakan SVM. Berdasarkan analisis hasil yang telah dilakukan, hasil klasifikasi terbaik oleh SVM memperoleh nilai akurasi sebesar 82,5% [17]. Kemudian penelitian lain melakukan analisis sentimen pengaruh jam kerja terhadap kesehatan mental generasi z. Hasil penelitian menunjukkan bahwa algoritma akurasi SVM mencapai 91% [29].

Penelitian lainnya melakukan klasifikasi tingkat stres dari data berbentuk teks dengan menggunakan algoritma SVM dan Random Forest. Model algoritma SVM dengan transformasi TF- IDF yang dibangun berhasil mendapatkan akurasi tertinggi sebesar 84% [30]. Selanjutnya terdapat penelitian lainnya tentang kemungkinan depresi dari postingan pada media sosial. Hasil pada penelitian mendapatkan akurasi paling tinggi adalah SVM yaitu 98.57% pada training data dan 95.56% pada test data [31].

Beberapa penelitian terdahulu menunjukkan bahwa nilai analisis sentimen terhadap stigma kesehatan mental menggunakan metode SVM memiliki hasil akurasi yang berbeda-beda. Seperti hasil penelitian ini yang mendapatkan model SVM sebagai model terbaik dengan akurasi yang cukup tinggi yaitu 86,11%. Nilai akurasi dari pemodelan tersebut dapat dipengaruhi oleh berbagai faktor, seperti kualitas data, teknik praproses yang digunakan dan cara pelabelan data. Selain itu penelitian ini memiliki beberapa perbedaan dari penelitian sebelumnya seperti data yang digunakan adalah data stigma kesehatan mental periode 2020-2023. Kemudian pelabelan data pada menggunakan kamus lexicon dengan memberikan polarity score. Penelitian ini juga menggunakan *count vectorizer* untuk melakukan ekstraksi fitur sebelum pemodelan.

5. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan dapat disimpulkan, pengumpulan data tweet mengenai stigma kesehatan mental dari periode 2020-2023 dengan *keyword* terkait menggunakan metode crawling menghasilkan 3.095 data. Setelah melalui praproses dan pelabelan data didapatkan 1.635 data (kelas negatif), 633 data (kelas positif) dan 208 data (kelas netral).

Hasil pengujian model SVM menunjukkan model dengan performa terbaik yaitu dengan akurasi 86.11%. Kemudian model Random Forest mendapatkan akurasi 82.55%. Sedangkan model Naive bayes mendapatkan akurasi sebesar 78,19%. Oleh karena itu, dapat disimpulkan bahwa SVM memiliki performa yang paling baik dalam mengklasifikasikan tweet yang mengandung stigma kesehatan mental.

Adapun saran untuk penelitian selanjutnya dapat melakukan penanganan ketidakseimbangan kelas. Karena kelas negatif biasanya paling mendominasi dibandingkan kelas positif dan netral, sehingga akan berpengaruh terahadap akurasi model. Selain itu dapat diterapkan *hyperparameter tuning* agar mendapatkan parameter terbaik untuk pembanguanan model sentimen analsis stigma kesehatan mental.

DAFTAR PUSTAKA

[1] D. Nur, N. S. Ibraya, and N. R. Marsuki, "Dampak Sosiologi Digital Terhadap Perubahan Sosial Budaya Pada Masyarakat Masa Depan Universitas Muhammadiyah Makassar interaksi sosial dan dinamika masyarakat . Berbagai perubahan khususnya dalam bidang," *J. Pendidik. Dan Ilmu Sos.*, vol. 2, no. 2, pp. 123–135, 2024.

[2] M. Fachriza and M. Munawar, "Analisis Sentimen Kalimat Depresi Pada Pengguna Twitter Dengan Naive Bayes, Support Vector Machine, Random Forest," *Komputek*, pp. 49–58, 2023, [Online]. Available: https://studentjournal.umpo.ac.id/index.php/komputek/article/view/2218

[3] R. R. Septa and B. Kusumasari, "Opini Publik Terkait Tren Isu Kesehatan: Analisis Konten pada Twitter dan Portal Berita di Yogyakarta," *J. Adm. Publik*, vol. 1, pp. 34–43, 2021, doi: 10.47753/pjap.v2i2.34.

[4] D. Ayuningtyas and M. Rayhani, "Analisis Situasi Kesehatan Mental Pada Masyarakat Di Indonesia Dan Strategi Penanggulangannya," *J. Ilmu Kesehat. Masy.*, vol. 9, no. 1, pp. 1–10, 2018.

[5] U. A. Annury, F. Yuliana, V. A. Z. Suhadi, and C. S. A. K. Karlina, "Dampak Self Diagnose Pada Kondisi Mental Health Mahasiswa Universitas Negeri Surabaya," *Jur. Ilmu Ilmu Sos. FISH Univ. Negeri Surabaya*, pp. 481–486, 2022.

[6] M. L. Wicaksono, R. Rusdah, and D. Apriana, "Sentiment Analysis Of Mental Health Using K-Nearest Neighbors On Social Media Twitter," *Bit (Fakultas Teknol. Inf. Univ. Budi Luhur)*, vol. 19, no. 2, p. 98, 2022, doi: 10.36080/bit.v19i2.2042.

[7] F. Anwar and P. Julia, "Analisis Strategi Pembinaan Kesehatan Mental Oleh Guru Pengasuh Sekolah Bersama si Aceh Besar Pada Masa Pandemi," *J. EDUKASI J. Bimbing. Konseling*, vol. 7, no. 1, pp. 64–83, 2021.

[8] 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.

[9] A. Ilham and W. Pramusinto, "Analisis Sentimen Masyarakat Terhadap Kesehatan Mental Pada Twitter Menggunakan Algoritme K-Nearest Neighbor," *Pros. Semin. Nas. Mhs. Fak. ...*, vol. 2, no. September, pp. 539–547, 2023, [Online]. Available: http://senafti.budiluhur.ac.id/index.php/senafti/article/view/792%0Ahttp://senafti.budiluhur.ac.id/index.php/senafti/article/download/792/527

[10] F. Andriani, "Fenomena Social Climber Melalui Twitwar," *J. Pustaka Komun.*, vol. 1, no. 2, pp. 349–360, 2018, [Online]. Available: https://journal.moestopo.ac.id/index.php/pustakom/article/view/713

[11] S. F. Pratama, R. Andrean, and A. Nugroho, "Analisis Sentimen Twitter Debat Calon Presiden Indonesia Menggunakan Metode Fined-Grained Sentiment Analysis," *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 4, no. 2, p. 39, 2019, doi: 10.31328/jointecs.v4i2.1004.

[12] E. R. Lidinillah, T. Rohana, and A. R. Juwita, "Analisis sentimen twitter terhadap steam menggunakan algoritma logistic regression dan support vector machine Steam sentiment analysis using logistic regression algorithm and support vector machine," vol. 10, pp. 154–164, 2023, doi: 10.37373/tekno.v10i2.440.

[13] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, p. 147, 2021, doi: 10.25126/jtiik.0813944.

[14] Y. Femilia Nugraini, R. Rohmat Saedudin, and R. Andreswari, "Implementasi Data Mining Dalam Kasus Mental Health Pada Sosial Media Twitter Menggunakan Metode Naive Bayes Implementation of Data Mining in the Case of Mental Health on Social Media Twitter Using Naive Bayes Method," *e-Proceeding Eng.*, vol. 8, no. 5, pp. 9260–9265, 2021, [Online]. Available: https://repository.telkomuniversity.ac.id/pustaka/files/170554/jurnal_eproc/implementasi-data-mining-dalam-kasus-mental-health-pada-sosial-media-twitter-menggunakan-metode-naive-bayes.pdf

[15] K. Yan, D. Arisandi, and T. Tony, "Analisis Sentimen Komentar Netizen Twitter Terhadap Kesehatan Mental Masyarakat Indonesia," *J. Ilmu Komput. dan Sist. Inf.*, vol. 10, no. 1, 2022, doi: 10.24912/jiksi.v10i1.17865.

[16] S. Mulyani and R. Novita, "Implementation of the Naive Bayes Classifier Algorithm for Classification of Community Sentiment About Depression on Youtube," *J. Tek. Inform.*, vol. 3, no. 5, pp. 1355–1361, 2022, doi: 10.20884/1.jutif.2022.3.5.374.

[17] F. Darmawan, M. Joe, Y. I. Kurniawan, and L. Afuan, "Analisis Sentimen Kemungkinan Depresi dan Kecemasan pada Twitter Menggunakan Support Vector Machine," *J. Eksplora Inform.*, vol. 13, no. 1, pp. 24–36, 2023, doi: 10.30864/eksplora.v13i1.854.

[18] Y. Fauziah, S. Saifullah, and A. S. Aribowo, "Design Text Mining for Anxiety Detection using Machine Learning based-on Social Media Data during COVID-19 pandemic," *Proceeding LPPM UPN "Veteran" Yogyakarta Conf. Ser. 2020 – Eng. Sci. Ser.*, vol. 1, no. 1, pp. 253–261, 2020, [Online]. Available: http://proceeding.rsfpress.com/index.php/ess/article/view/117

[19] S. Algifari Rismawan and Y. Syahidin, "Implementasi Website Berita Online Menggunakan Metode Crawling Data Dengan Bahasa Pemrograman Python," *J. Tek. Inform. dan Sist. Inf.*, vol. 10, no. 3, pp. 167–178, 2023, [Online]. Available: https://doi.org/10.35957/jatisi.v10i3.4902

[20] A. Nofandi, N. Setiawan, and D. Brata, "Analisis Sentimen Ulasan Pelanggan dengan Metode Support Vector Machine (SVM) untuk Peningkatan Kualitas Layanan pada Restoran Warung Wareg," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 1, pp. 458–466, 2023, [Online]. Available: http://j-ptiik.ub.ac.id

[21] A. R. Ismail and R. B. F. Hakim, "Implementasi Lexicon Based Untuk Analisis Sentimen Dalam Menentukan Rekomendasi Pantai Di DI Yogyakarta Berdasarkan Data Twitter," *Emerg. Stat. Data Sci. J.*, vol. 1, no. 1, pp. 37–46, 2023, doi: 10.20885/esds.vol1.iss.1.art5.

[22] N. Yusliani, A. Yuhafiz, M. D. Marieska, and A. S. Utami, "Analisis Sentimen di Twitter Menggunakan Algoritma Artificial Neural Network," *J. Jupiter*, vol. 15, no. 1, pp. 725–731, 2023.

[23] R. Vincent *et al.*, "Perbandingan Klasifikasi Naive Bayes Dan Support Vector Machine Dalam Analisis Sentimen Dengan Multiclass Di Twitter," *J. Mhs. Tek. Inform.*, vol. 7, no. 4, pp. 2496–2505, 2023.

[24] S. Amaliah, M. Nusrang, and A. Aswi, "Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 4, no. 3, pp. 121–127, 2022, doi: 10.35580/variansiunm31.

[25] J. Melvin and A. Soraya, "Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit," *J. Ris. Rumpun Mat. dan Ilmu Pengetah. Alam*, vol. 2, no. 2, pp. 87–103, 2023.

[26] I. Kurniawan, D. C. P. Buani, A. Abdussomad, W. Apriliah, and R. A. Saputra, "Implementasi Algoritma Random Forest Untuk Menentukan Penerima Bantuan Raskin," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 2, pp. 421–428, 2023, doi: 10.25126/jtiik.20231026225.

[27] Rayuwati, Husna Gemasih, and Irma Nizar, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid," *Jural Ris. Rumpun Ilmu Tek.*, vol. 1, no. 1, pp. 38–46, 2022, doi: 10.55606/jurritek.v1i1.127.

[28] R. Nurhidayat and K. E. Dewi, "Penerapan Algoritma K-Nearest Neighbor Dan Fitur Ekstraksi N-Gram Dalam Analisis Sentimen Berbasis Aspek," *Komputa J. Ilm. Komput. dan Inform.*, vol. 12, no. 1, pp. 91–100, 2023, doi: 10.34010/komputa.v12i1.9458.

[29] M. Daffa, A. Fahreza, A. Luthfiarta, M. Rafid, M. Indrawan, and A. Nugraha, "Analisis Sentimen : Pengaruh Jam Kerja Terhadap Kesehatan Mental Generasi Z," *J. Appl. Comput. Sci. Technol.*, vol. 5, no. 1, pp. 16–25, 2024.

[30] N. Fathirachman Mahing, A. Lazuardi Gunawan, A. Foresta Azhar Zen, F. Abdurrachman Bachtiar, and S. Agung Wicaksono, "Klasifikasi Tingkat Stress dari Data Berbentuk Teks dengan Menggunakan Algoritma Support Vector Machine (SVM) dan Random Forest," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 7, pp. 1527–1536, 2023, doi: 10.25126/jtiik.1078010.

[31] S. Mutmainah, "Kemungkinan Depresi Dari Postingan Pada Sosial Media," *J. Sains, Nalar, dan Apl. Teknol. Inf.*, vol. 1, no. 2, pp. 17–23, 2022, doi: 10.20885/snati.v1i2.11.