Jurnal Komputer dan Informatika
Vol 18 No 2, Oktober 2023: hlm 133 - 140

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi

KLASIFIKASI DATA TEKS UNTUK MENDETEKSI EMOSI PENGGUNA TWITTER MENGGUNAKAN MACHINE LEARNING

Yosia A. Ishak
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara
Jln. Letjen S. Parman No. 1, Jakarta, 11440, Indonesia
e-mail: *535210037@stu.untar.ac.id*

ABSTRAK

Pendeteksian emosi pada platform media sosial seperti twitter merupakan sebuah tantangan di dunia pemrosesan data. Berbeda dengan kehidupan nyata, deteksi emosi pada teks adalah tugas yang kompleks dikarenakan kurangnya konteks atau beberapa kalimat tidak disertakan dengan konteks yang cukup untuk memahami emosi yang terkandung didalamnya. Penelitian ini bertujuan untuk membuat model yang mumpuni dalam memenuhi kebutuhan klasifikasi emosi berdasarkan teks pada aplikasi Twitter. Pada penelitian ini, dilakukan perbandingan antara algoritma Naïve Bayes dan LinearSVC, dan didapatkan hasil akhir dimana algoritma LinearSVC mencapai akurasi terbaik dengan 66%, sedangkan algoritma Naïve Bayes mencapai akurasi 57%. Dengan dilakukannya penelitian ini, diharapkan dapat memberikan kontribusi dalam mengembangkan pemahaman terhadap analisis sentimen di media sosial dan membuka peluang untuk aplikasi lebih lanjut, seperti pemantauan opini publik dan analisis tren emosional secara *real-time.*

Kata kunci: Twitter, Klasifikasi, Teks, Naïve Bayes, LinearSVC.

ABSTRACT

Emotion detection on social media platforms such as Twitter is a challenge in the world of data processing. In contrast to real life, detecting emotions in text is a complex task due to a lack of context or some sentences are not provided with sufficient context to understand the emotions contained therein. This research aims to create a model that is capable of meeting the needs of text-based emotion classification in the Twitter application. In this research, a comparison was carried out between the Naïve Bayes and LinearSVC algorithms, and the final results were obtained where the LinearSVC algorithm achieved the best accuracy with 66%, while the Naïve Bayes algorithm achieved 57% accuracy. By conducting this research, it is hoped that it can contribute to developing understanding of sentiment analysis on social media and open up opportunities for further applications, such as monitoring public opinion and analyzing emotional trends in real-time.

Keywords: *Twitter, Classification, Text, Naïve Bayes, LinearSVC.*

1. PENDAHULUAN

Media sosial adalah alat di Internet yang memungkinkan pengguna untuk mewakili diri mereka sendiri dan secara virtual berinteraksi, berkolaborasi, berbagi, berkomunikasi dengan pengguna lain dan membentuk ikatan sosial [1]. Saat ini perkembangan sosial media semakin masif, aktivitas manusia di sosial media juga semakin beragam, mulai dari untuk komunikasi, berjualan, dan aktivitas lainnya [2]. Beberapa aktivitas yang dapat dilakukan di media sosial, misalnya yaitu melakukan komunikasi atau interaksi hingga memberikan informasi atau konten berupa tulisan, foto dan video [3].

Di era digital saat ini, media sosial khususnya Twitter telah menjadi platform penting bagi pengguna untuk berbagi pemikiran, pendapat, dan emosi. Twitter merupakan situs jejaring sosial yang diluncurkan pada 2006 dan kini setidaknya memiliki 100 juta pengguna aktif setiap hari, dan 500 juta tweet yang dikirim setiap hari [4]. Menurut situs databoks, pengguna aktif aplikasi Twitter di Indonesia menempati urutan ke-6 terbanyak di dunia [5].

Munculnya sejumlah besar data tekstual yang dihasilkan oleh pengguna Twitter memberikan peluang besar untuk menganalisis dan memahami emosi dan sentimen yang terkandung dalam setiap postingan dengan melakukan klasifikasi data teks. Klasifikasi teks adalah proses pengelompokan data kedalam kelas yang telah ditentukan sebelumnya, untuk bisa digunakan memprediksi kelas dari data-data yang kelas belum diketahui [6]. Teks adalah rangkaian kata atau kalimat yang memiliki struktur dan tata bahasa tertentu serta bisa disusun secara lisan maupun tulisan. Tujuannya, untuk menyampaikan informasi, menjelaskan sesuatu, atau mengungkapkan makna [7].

Tujuan dari penelitian ini adalah membandingkan kekuatan pembelajaran mesin untuk mengembangkan model klasifikasi menggunakan algoritma Naïve Bayes dan LinearSVC yang dapat mengenali emosi pengguna Twitter dari data teks yang diperoleh. [8]. Algoritma LinearSVC adalah algoritma yang menerapkan fungsi kernel linier untuk melakukan klasifikasi dan bekerja dengan baik dengan jumlah sampel yang besar [9].

2. TINJAUAN LITERATUR

Adapun penelitian terkait klasifikasi teks:

a) *Systematic Literature Revie*w: Metode *Machine Learning* dalam Klasifikasi Emosi pada Data Tekstual

Penelitian ini bertujuan untuk mengidentifikasi serta menganalisis dataset, metode, dan metrik evaluasi yang digunakan dalam penelitian mengenai klasifikasi emosi pada data tekstual dari tahun 2013 sampai tahun 2022. Berdasarkan desain inklusi dan eksklusi dalam pemilihan literatur, didapatkan sebanyak 50 literatur yang akan digunakan dalam ekstraksi dan sintesis data. Hasil analisis menunjukkan bahwa dari 50 literatur, 36 literatur menggunakan dataset publik dan 14 literatur menggunakan dataset privat. Pada metode dalam pengembangan model, SVM dan Naive Bayes merupakan model yang paling populer diantara lainnya. Dalam melakukan evaluasi model, metrik F-measure atau F1-score adalah metrik yang paling banyak digunakan dibandingkan dengan matrik lainnya. [10].

b) Klasifikasi Emosi pada Teks dengan Menggunaka*n Metode Deep Learning*

Penelitian ini membahas klasifikasi emosi pada teks dengan menggunakan *metode deep learning*. Model yang digunakan adalah BERT (*Bidirectional Encoder Representations from Transformers*) dan dilakukan pada data opini. Hasil pengujian menunjukkan bahwa model mampu mendeteksi emosi pada teks dengan tepat [11].

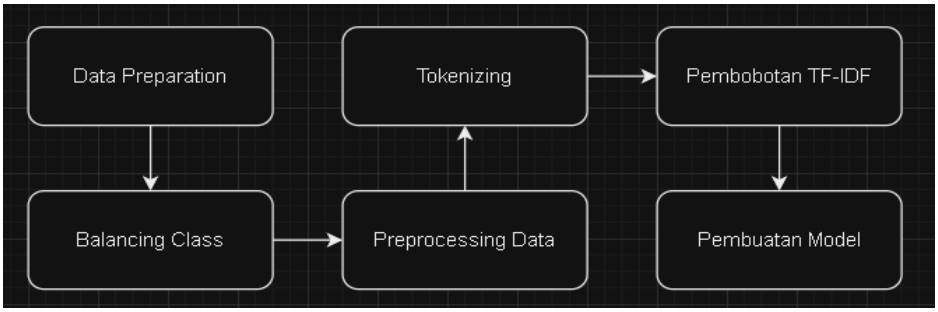
c) Implementasi Machine Learning untuk Mendeteksi Unsur Depresi pada Tweet Menggunakan *Metode Naïve*

Jurnal ini membahas tentang penggunaan metode Naïve Bayes untuk mendeteksi unsur depresi pada tweet. Hasil pengujian menunjukkan bahwa sistem deteksi unsur depresi pada tweet dengan metode *Naïve Bayes* dapat mendeteksi sebuah postingan tweet mengandung unsur depresi atau tidak dengan nilai akurasi sebesar 74% pada pengujian rasio data, dan 73% pada pengujian validasi [12].

3. METODE PENELITIAN

Penelitian ini akan memiliki beberapa langkah-langkah yang diperlukan, rincian tahapan penelitian adalah sebagai berikut:

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi



Gambar 1. Tahapan Penelitian

1.1 Sumber Dataset

Dataset didapatkan dari laman *Kaggle* yang merupakan sebuah situs daring yang menyediakan berbagai sumber daya dan kompetisi dalam bidang ilmu data dan coding [13] . Total data berjumlah 40.000 data, terdapat 13 label emosi yaitu *anger, boredom, empty, enthusiasm, fun, happiness, hate, love, neutral, relief, sadness, surprise, worry,* dan terdapat 3 kolom yaitu *tweet_id, sentiment,* dan *content*. Semua tipe data adalah string kecuali untuk kolom *tweet_id* yang bertipe integer. Terdapat singkatan kata atau kata informal seperti *peeps* yang merupakan kata lain dari *people.* Perlu diingat bahwa dataset ini mengalami *class imbalance* atau ketidaksamaan jumlah data pada setiap label sebesar <21.32% [14].

1.2 *Preprocessing* Data

Preprocessing data adalah proses persiapan dan transformasi data mentah menjadi format yang lebih terstruktur dan siap analisis. Proses ini sangat penting dalam analisis data dan pembelajaran mesin karena membantu dalam menghasilkan data yang berkualitas dan dapat diandalkan untuk keperluan analisis lebih lanjut [15].

1.2.1 *Case Folding*

Case folding adalah salah satu bentuk text preprocessing yang paling sederhana dan efektif meskipun sering diabaikan. Tujuan dari case folding untuk mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf ‘a’ sampai ‘z’ yang diterima [16].

1.2.2 *Tokenizing*

Tokenizing adalah operasi memisahkan teks menjadi potongan-potongan berupa token, bisa berupa potongan huruf, kata, atau kalimat, sebelum dianalisis lebih lanjut [17]. Pada dasarnya tujuan dari tokenisasi ini adalah untuk memecahkan sebuah kalimat ke dalam kepingan-kepingan kata. Misalnya kalimat “Ini baju saya” akan menjadi “ini, baju, saya” [18].

1.2.3 Filtering

Filtering adalah salah satu metode *preprocessing* data yang digunakan untuk mengambil kata-kata yang penting dari hasil token tadi. Kata umum yang biasanya muncul dan tidak memiliki makna disebut dengan stopword [19]. Misalnya terdapat kalimat “manajemen pengetahuan adalah sebuah konsep baru di dunia bisnis”, hasil dari filtering kalimat tersebut adalah “manajemen, pengetahuan, konsep, baru, dunia, bisnis”.

1.2.4Stemming

Stemming merupakan suatu proses untuk menemukan kata dasar dari sebuah kata [20]. Misalnya kata “membela” menjadi “bela”, dan “dibandingkan” menjadi “banding”.

1.3 Ekstraksi Fitur (TF-IDF**)**

TF-IDF atau *Term Frequency-Inverse Document Frequency* (merupakan metode yang digunakan untuk menentukan nilai frekuensi sebuah kata di dalam sebuah dokumen atau artikel dan

juga frekuensi di dalam banyak dokumen. *Term frequency* (TF) merupakan frekuensi kemunculan kata (t) pada kalimat (d). *Document frequency* (DF) adalah banyaknya kalimat dimana suatu kata (t) muncul [21]. Perhitungan ini menentukan seberapa relevan sebuah kata di dalam sebuah dokumen. TF-IDF dapat dirumuskan sebagai berikut:

$$TF - IDF(t_k, d_j) = TF(t_k, d_j) * (IDF(t_k,) \quad (1)$$

dimana :
d_j = Dokumen ke-j
t_k = Term ke-k

$$TF(t_k, d_j) = f(t_k, d_j) \quad (2)$$

dimana :
TF = Jumlah frekuensi term
f = Jumlah frekuensi kemunculan
d_j = Dokumen ke-j
t_k = Term ke-k

$$IDF(t_k) = \frac{1}{df(t)} \quad (3)$$

dimana :
IDF = Jumlah frekuensi term
N = Jumlah frekuensi kemunculan
df = Jumlah kemunculan dokumen
d_j = Dokumen ke-j
t_k = Term ke-k

1.4 Pelatihan Model *Naïve Bayes*

Algoritma Naive Bayes adalah algoritma klasifikasi atau prediksi peluang di masa yang akan datang berdasarkan pengalaman yang sudah terjadi sebelumnya menggunakan probabilitas yang dicetuskan oleh ilmuwan inggris yaitu Thomas Bayes [22]. Rumus umum dasar algoritma Naïve Bayes ditunjukkan pada rumus 4 [23].

$$P(X) = \frac{P(H) \cdot P(H)}{P(X)} \quad (4)$$

dimana:
X = data tanpa kelas
H = Hipotesis data X merupakan suatu kelas spesifik
P(X) = Probabilitas hipotesis X
P(H) = Probabilitas hipotesis H
P(X|H) = Probabilitas X berdasarkan kondisi pada hipotesis H

P(H|X) = Probabilitas hipotesis H berdasarkan kondisi X

1.5 Pelatihan Model LinearSVC

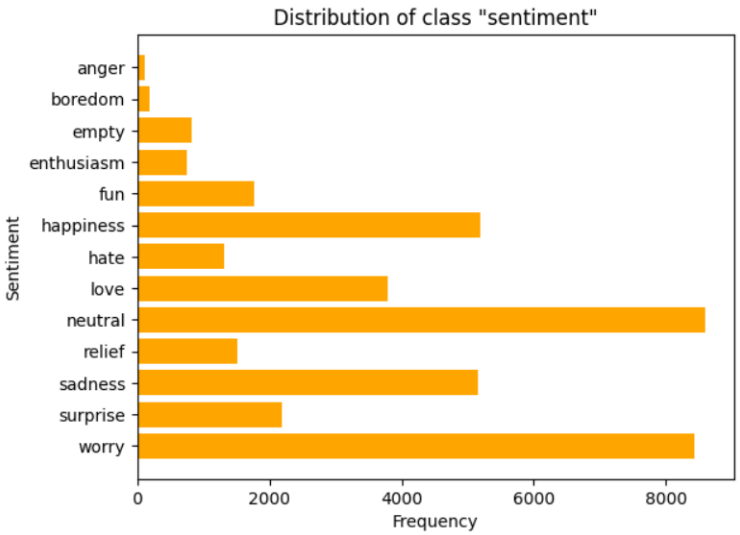
LinearSVC (*Linear Support Vector Classificatio*n) adalah algoritma yang digunakan untuk pemodelan klasifikasi. Algoritma ini termasuk dalam keluarga *Support Vector Machine* (SVM) dan khusus digunakan untuk masalah klasifikasi dua kelas (klasifikasi biner). Algoritma LinearSVC bekerja dengan mencari hyperplane terbaik yang memisahkan dua kelas dengan batas keputusan terbesar. Hyperplane ini dipilih sedemikian rupa sehingga jarak antara hyperplane dan vertex (data yang paling dekat dengan batas keputusan) dimaksimalkan.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi

Penting untuk dicatat bahwa LinearSVC berfungsi dengan baik untuk data yang dapat dipisahkan secara linier atau data dapat dipisahkan dengan garis lurus. Jika data tidak dapat dipisahkan secara linier, atau jika hubungan antara fitur dan target bersifat nonlinier, mungkin akan memerlukan pendekatan lain, seperti menggunakan kernel nonlinier dengan SVM.

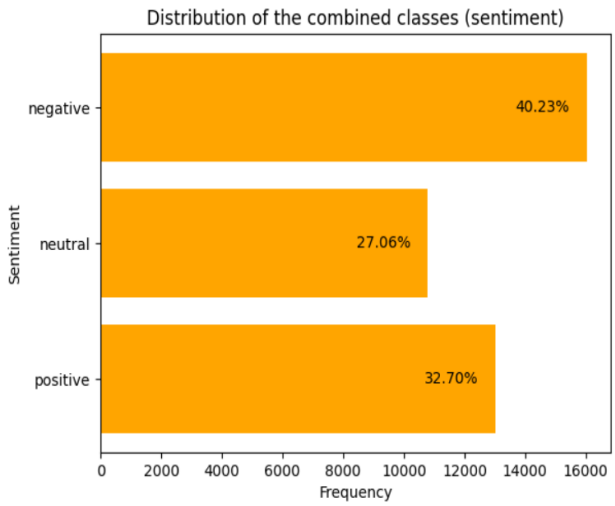
4. HASIL DAN PEMBAHASAN

Setelah melewati proses processing yang terdiri dari case folding, tokenizing, filtering, stemming, pembobotan TF-IDF, berikut merupakan plot distribusi dari kelas "sentiment":



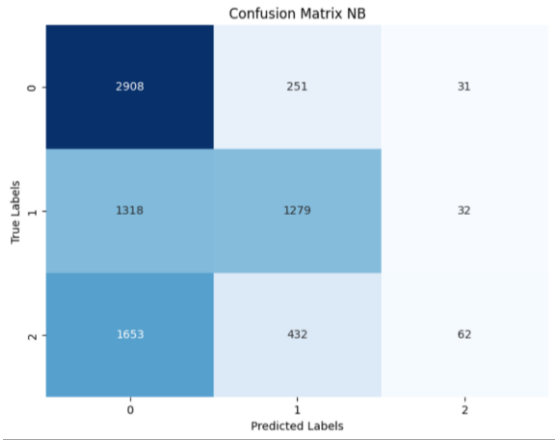
Gambar 2. Plot Distribusi Kelas "Sentiment"

Terjadi fenomena *data imbalance* atau ketidakseimbangan data, untuk mengatasi hal tersebut 13 label dikelompokkan menjadi 3 kelas utama, yaitu *negative*, *positive*, dan *neutral.* Pembagian dilakukan dengan kolom *empty*, *sadness*, *worry*, *hate*, *boredom*, *anger* dikelompokkan kedalam kelas *negative*, kolom *enthusiasm*, *love*, *fun*, *happiness*, *relief* dikelompokkan kedalam kelas *positive*, dan kolom *surprise* serta *neutral* dikelompokkan kedalam kelas *neutral.* Untuk catatan, kolom *surprise* dimasukkan kedalam kelas *neutral*, hal ini dikarenakan sentiment tersebut dapat memicu sesuatu yang positif ataupun sesuatu yang negatif, seperti terkejut mendapatkan hadiah dan terkejut melihat sesuatu yang menakutkan.

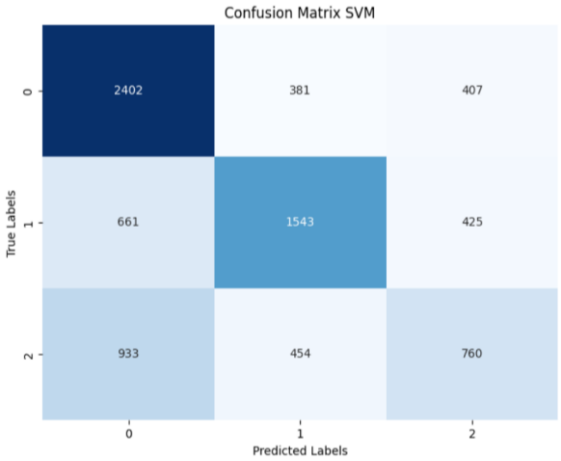


Gambar 3. Plot Distribusi Pengelompokan Kelas "Sentiment"

Data kemudian dibagi menjadi data latih dan data uji, kemudian dibuat model dengan algoritma *Naïve Bayes* dan algoritma LinearSVC. Metrik evaluasi terdiri dari precision, recall, *f1-score*, dan *support*. Berikut merupakan perbandingan heatmap dari confusion matrix dan *classification report* kedua algoritma tersebut:



Gambar 4. Heatmap Confusion Matrix Algoritma Naïve Bayes

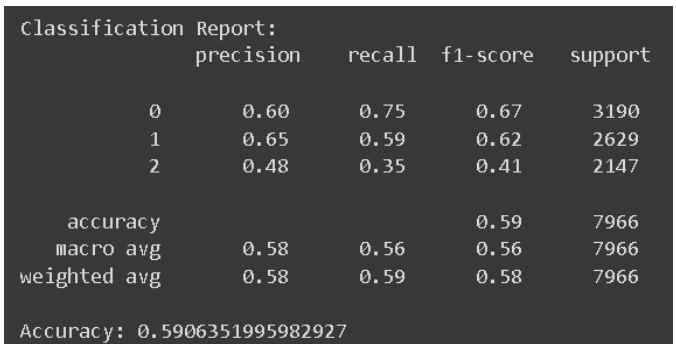


Gambar 5. Heatmap Confusion Matrix Algoritma Linear SVC

```
Classification Report (Cross-Validation):
              precision    recall  f1-score   support

           0       0.49      0.92      0.64     12834
           1       0.63      0.45      0.53     10395
           2       0.53      0.02      0.05      8632

    accuracy                           0.53     31861
   macro avg       0.55      0.47      0.41     31861
weighted avg       0.55      0.53      0.44     31861

Accuracy (Cross-Validation): 0.5254386240231004
Classification Report (Test Data):
              precision    recall  f1-score   support

           0       0.49      0.91      0.64      3190
           1       0.65      0.49      0.56      2629
           2       0.50      0.03      0.05      2147

    accuracy                           0.53      7966
   macro avg       0.55      0.48      0.42      7966
weighted avg       0.55      0.53      0.46      7966

Accuracy (Test Data): 0.5333919156414763
```

Gambar 6. Classification Report Algoritma Naïve Bayes

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi



```
Classification Report:
              precision    recall  f1-score   support

           0       0.60      0.75      0.67      3190
           1       0.65      0.59      0.62      2629
           2       0.48      0.35      0.41      2147

    accuracy                           0.59      7966
   macro avg       0.58      0.56      0.56      7966
weighted avg       0.58      0.59      0.58      7966

Accuracy: 0.5906351995982927
```

Gambar 7. Classification Report Algoritma Linear SVC

Tabel 1. Perbandingan Akurasi

Algoritma	Akurasi
Naïve Bayes	0.53
LinearSVC	0.59

5. KESIMPULAN

Melalui analisis dan eksperimen yang dilakukan serta perbandingan dari kedua algoritma, didapatkan hasil bahwa algoritma LinearSVC mencapai akurasi sebesar 59%, sedangkan akurasi algoritma Naïve Bayes sebesar 53%, sehingga dapat disimpulkan algoritma LinearSVC merupakan algoritma yang lebih baik daripada algoritma Naïve Bayes dalam menyelesaikan masalah ini.

Namun, tentu masih terdapat ruang untuk pengembangan lebih lanjut. Perlu dilakukan penelitian lanjutan guna meningkatkan kinerja dan akurasi model, mengingat dinamika bahasa dan perkembangan tren yang terus berubah di platform media sosial. Untuk rencana pengembangan selanjutnya, diharapkan dapat dilakukan penelitian yang lebih mendalam dengan menggunakan algoritma klasifikasi lainnya untuk mencari kemungkinan kinerja algoritma yang lebih baik daripada algoritma yang telah diuji.

DAFTAR PUSAKA

[1]. F. Yusuf, H. Rahman, S. Rahmi dan A. Lismayani, “Pemanfaatan Media Sosial Sebagai Sarana Komunikasi, Informasi, Dan Dokumentasi: Pendidikan Di Majelis Taklim Annur Sejahtera,” Jurnal Hasil-Hasil Pengabdian dan Pemberdayaan Masyarakat, vol. 2, no. 1, p. 3, 2023.

[2] N. A. Siti, “Menilik Sejarah Media Sosial, Manfaat, dan Contohnya,” katadata, 1 April 2022. [Online]. Available: https://katadata.co.id/sitinuraeni/digital/6246823429ac2/menilik-sejarah-media-sosial-manfaat-dan-contohnya. [Diakses 4 December 2023].

[3] Nandy, “Pengertian Media Sosial, Sejarah, Fungsi, Jenis, Manfaat, dan Perkembangannya,” Gramedia, [Online]. Available: https://www.gramedia.com/literasi/pengertian-media-sosial/. [Diakses 4 December 2023].

[4] N. R. Aida dan S. Hardiyanto, “Mengenal Apa Itu Twitter dan Mengapa Orang Menggunakannya?,” Kompas, 24 March 2022. [Online]. Available: https://www.kompas.com/tren/read/2022/03/24/200500665/mengenal-apa-itu-twitter-dan-mengapa-orang-menggunakannya?page=all. [Diakses 4 December 2023].

[5] O. S. Y. Prakasa dan K. M. Lhaksamana, “Klasifikasi Teks Dengan Menggunakan Algoritma K-Nearest Pada Kasus Kinerja Pemerintah Di Twitter,” *e-Proceeding of Engineering,* vol. 5, no. 3, p. 8238, 2018.

[6] anugrah dwi, “Pengertian Teks Beserta Jenisnya Lengkap,” Kampus Pascasarjana UMSU, 27 January 2023. [Online]. Available: https://pascasarjana.umsu.ac.id/pengertian-teks-beserta-jenisnya-lengkap/. [Diakses 4 December 2023].

[7] R. Tineges, "Mengenal Naive Bayes Sebagai Salah Satu Algoritma Data Science," DQLab, 23 May 2022. [Online]. Available: https://dqlab.id/mengenal-naive-bayes-sebagai-salah-satu-algoritma-data-science. [Diakses 5 December 2023].
[8] Classification Example with Linear SVC in Python," DataTechNotes, 7 January 2020. [Online]. Available: https://www.datatechnotes.com/2020/07/classification-example-with-linearsvm-in-python.html. [Diakses 4 December 2023].
[9] C. Pramartha dan P. W. A. Wibawa, "Systematic Literature Review: Metode Machine Learning dalam Klasifikasi Emosi pada Data Tekstual," *Jurnal Sistem Informasi dan Komputer (SISFOKOM),* vol. 12, no. 03, 2023.
[10] D. H. Fudholi dan A. T. B. W, "Klasifikasi Emosi Pada Teks Dengan Menggunakan Metode Deep Learning," *Jurnal Ilmiah Indonesia,* vol. 6, no. 1, 2021
[11] M. M. Nurrochman, A. L. Prasasti dan R. A. Nugrahaeni, "Implementasi Machine Learning Untuk Mendeteksi Unsur Depresi Pada Tweet Menggunakan Metode Naïve Bayes," *e-Proceeding of Engineering,* vol. 8, no. 5, 2021.
[12] Kaggle: Tempat Belajar Data Science," Course-Net, 27 November 2023. [Online]. Available: https://course-net.com/blog/kaggle-tempat-belajar-data-science/. [Diakses 4 December 2023].
[13] P. Gupta, "Emotion Detection from Text," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text. [Diakses 4 December 2023].
[14] N. L. Kamila, "Mengenal Apa Itu Tahap Preprocessing Data," dibimbing, 25 November 2023. [Online]. Available: https://dibimbing.id/blog/detail/mengenal-apa-itu-tahap-preprocessing-data. [Diakses 4 December 2023].
[15] K. S. Nugroho, "Dasar Text Preprocessing dengan Python," Medium, 18 June 2019. [Online]. Available: https://ksnugroho.medium.com/dasar-text-preprocessing-dengan-python-a4fa52608ffe. [Diakses 4 December 2023].
[16] A. Rizki, "#BelajarPython 9: Operasi 'Tokenizing' pada Teks Berbahasa Indonesia," 16 August 2020. [Online]. Available: https://adityarizki.net/belajarpython-9-operasi-tokenizing-pada-teks-berbahasa-indonesia/. [Diakses 4 December 2023].
[17] Thalib, "NLP Preprocessing : Teknik Tokenisasi Untuk Memecah Kalimat menjadi Kata-Kata Pada Python," Medium, 1 November 2019. [Online]. Available: https://medium.com/@irfandy.thalib/teknik-tokenisasi-untuk-memecah-kalimat-menjadi-kata-kata-pada-python-12f799b74d49. [Diakses 4 November 2023].
[18] R. Tineges, "Tahapan Text Preprocessing dalam Teknik Pengolahan Data," DQLab, 17 June 2021. [Online]. Available: https://dqlab.id/tahapan-text-preprocessing-dalam-teknik-pengolahan-data. [Diakses 4 December 2023
[19] M. U. Albab, Y. Karuniawati dan M. N. Fawaiq, "Optimization of the Stemming Technique on Text preprocessing President 3 Periods Topic," *Jurnal Transformatika,* vol. 20, no. 2, p. 1, 2023.
[20] E. T. Wijaya, "PERANCANGAN INFORMATION RETRIEVAL (IR) BERBASIS TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF) UNTUK PERINGKASAN TEKS TUGAS KHUSUS BERBAHASA INDONESIA," *Jurnal Ilmiah Teknologi dan Informasi ASIA,* vol. 7, no. 1, p. 82, 2013.
[21 Syarli dan A. A. Muin, "Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru Perguruan Tinggi)," *Jurnal Ilmiah Ilmu Komputer,* vol. 2, no. 1, p. 22, 2016.
[22] A. Imran, M. Akbar, A. Desiani dan I. Irmeilyana, "Implementasi Algoritma Naive Bayes dan Support Vector Machine (SVM) Pada Klasifikasi Penyakit Kardiovaskular," *Jurnal Teknik Elektro dan Komputasi (ELKOM),* vol. 4, no. 2, p. 209, 2022.