



Text Classification for Detecting Depression and Anxiety among Twitter Users based on Machine Learning

Klasifikasi Teks untuk Mendeteksi Depresi dan Kecemasan Pada Pengguna Twitter Berbasis Machine Learning

**Kharisma Rahayu¹, Vindi Fitria², Dhini Septhya³,
Rahmaddeni^{4*}, Lusiana Efrizoni⁵**

^{1,2,3,4,5}Program Studi Teknik Informatika, STMIK Amik Riau, Pekanbaru, Riau

E-Mail: ¹2010031802129@sar.ac.id, ²2010031802085@sar.ac.id, ³2010031802093@sar.ac.id,
⁴rahmaddeni@sar.ac.id, ⁵lusiana@stmik-amik-riau.co.id

Received Jun 4th 2023; Revised Jul 14th 2023; Accepted Aug 20th 2023
Corresponding Author: Rahmaddeni

Abstract

Depression is a mental disorder that affects a person's well-being and quality of life. Social influences cause people with depression and anxiety disorders to ignore those around them. So they turn to social media like Twitter for support. Depressed people using Twitter are difficult to identify based on tweets. Therefore, text classification modeling for depression sufferers is needed to find out how many twitter users experience depression and anxiety. Text classification is a method to group similar data into appropriate groups. The algorithms used in making patterns of depression sufferers are Decision Tree, Random Forest, Naive Bayes, and K-Nearest Neighbor (KNN) algorithms. The use of the four algorithms aims to see the performance of the best algorithm. The results of the experiments conducted obtained that the Random Forest algorithm on splitting data 80:20 has better performance, with an accuracy value of 0.957 or 96%. The results of this study can be used by users to find out sufferers of depression and anxiety.

Keyword: Classification, Decision Tree, Depression, K-Nearest Neighbor, Naive Bayes, Random Forest

Abstrak

Depresi adalah gangguan mental yang mempengaruhi kesejahteraan dan kualitas hidup seseorang. Pengaruh sosial menyebabkan penderita depresi dan gangguan kecemasan mengabaikan orang-orang di sekitarnya. Jadi mereka beralih ke media sosial seperti *Twitter* untuk mendapatkan dukungan. Penderita depresi pengguna *Twitter* sulit untuk diidentifikasi berdasarkan tweet. Oleh karena itu pemodelan klasifikasi teks untuk penderita depresi sangat diperlukan agar mengetahui seberapa banyak pengguna *twitter* yang mengalami depresi dan kecemasan. Klasifikasi teks merupakan metode untuk mengelompokkan data yang sejenis ke dalam kelompok yang sesuai. Algoritma yang digunakan dalam membuat pola penderita depresi adalah algoritma *Decision Tree*, *Random Forest*, *Naive Bayes*, dan *K-Nearest Neighbor (KNN)*. Penggunaan keempat algoritma bertujuan untuk melihat kinerja algoritma terbaik. Hasil percobaan yang dilakukan diperoleh bahwa algoritma *Random Forest* pada splitting data 80:20 memiliki kinerja yang lebih baik, dengan nilai akurasi sebesar 0.957 atau 96%. Hasil penelitian ini dapat digunakan oleh pengguna untuk mengetahui penderita depresi dan kecemasan.

Kata Kunci: Decision Tree, Depresi, Klasifikasi, K-Nearest Neighbor, Naive Bayes, Random Forest

1. PENDAHULUAN

Media sosial saat ini menjadi wadah dimana pengguna dapat mengungkapkan perasaan, emosi, maupun opini mereka tentang topik atau permasalahan apapun secara daring. *Twitter* menjadi wadah publik yang digunakan pengguna untuk mengekspresikan diri, dari media sosial ini kita mendapatkan informasi tentang banyak hal dari *tweet* para pengguna *Twitter* [1]. Informasi dalam bentuk apapun dapat menyebar dengan cepat dan mudah, mempengaruhi pandangan, cara hidup, dan budaya suatu bangsa [2], *Twitter* merupakan media sosial berbentuk *microblog* atau *short pieces* yang mencapai 280 karakter, karena jumlah *tweet* terbatas atau dibatasi [3].

Depresi didefinisikan sebagai perasaan sakit atau patah semangat yang ditandai adanya keresahan berlebih, termasuk kehilangan dan ketegangan, dan beberapa gejala seperti kesulitan tidur dan hilangnya nafsu makan. Selain itu, depresi adalah salah satu yang menyebabkan kecacatan paling umum di seluruh dunia, termasuk bunuh diri, dengan perkiraan 300 juta orang depresi di seluruh dunia setiap tahun. Depresi didiagnosa secara pribadi menggunakan kriteria depresi klinis. Namun, 70% pasien depresi tahap awal tidak ingin melakukan konsultasi dengan dokter karena dapat memperburuk kondisi. Disaat yang bersamaan, orang terus mengandalkan jejaring sosial kemudian mengungkapkan perasaannya melalui jejaring sosial [4]. Hal ini menunjukkan bahwa media sosial dapat menjadi sumber lain untuk membantu profesional kesehatan mental menarik kesimpulan dan menemukan petunjuk tentang gangguan mental seperti depresi. Klasifikasi yaitu proses mengelompokkan sesuatu yang memiliki karakteristik yang mirip kedalam beberapa kelas. Pada umumnya pengklasifikasian dokumen diwakili oleh kalimat-kalimat penting dengan menentukan ciri-ciri atau karakteristik [5]. Salah satu metode klasifikasi yaitu *Decision Tree*, *Random Forest*, *Naive Bayes*, dan *K-Nearest Neighbor*.

Decision Tree adalah algoritma populer dan sangat efektif dengan melakukan pengklasifikasian dan prediksi. Algoritma *Decision Tree* dapat merepresentasikan ketentuan dari banyaknya fakta ke dalam bentuk pohon keputusan. Pohon keputusan adalah struktur yang membagi sejumlah besar data menjadi sejumlah kecil data. Atribut kelas berfungsi sebagai representasi untuk simpul daun pohon keputusan. *Node* yang tidak ada termasuk *node* internal yang dihasilkan oleh kondisi uji atribut pada beberapa *record* dengan berbagai karakteristik dan *node* akhir yang terdiri dari akar [5]. *Random Forest* mempunyai beberapa kelebihan yang dapat menaikkan akurasi hasil data yang hilang dan *resiting outliers*, serta menyimpan data yang efisien. Selain itu, *random forest* memiliki proses fitur seleksi yang dapat memperoleh fitur terbaik untuk menaikkan performa dari model klasifikasi [6]. Untuk meningkatkan kinerja model dari ke 4 algoritma (i.e., *Decision Tree*, *Random Forest*, *Naive Bayes*, dan *K-Nearest Neighbor*), peneliti menambahkan *feature extraction* untuk meningkatkan kinerja model.

Dari penelitian sebelumnya oleh Alec Go et al., yang mengklasifikasikan sentimen di *Twitter* menggunakan algoritma *Maximum Entropy*, *Naive Bayes*, dan *Support Vector Machine*. Akurasi hasil penelitian lebih dari 80% [9]. Studi lain mendeteksi depresi pada unggahan di media sosial *Reddit*. Studi ini menggunakan teknik pemrosesan bahasa alami bersama dengan metode pembelajaran mesin kemudian melatih data dan efektivitas teknik klasifikasi *Support Vector Machine (SVM)* dievaluasi dan mencapai angka 80%. Ketepatan studi ini juga menggunakan gabungan fitur tersukses (LIWC, LDA, dan bigram) dengan pengklasifikasi multilayer perceptron (MLP), yang memberikan performa tertinggi untuk mendeteksi depresi dan akurasi mencapai 91% [10]. Penelitian lain menggunakan *Naive Bayes (NB)* dan *Support Vector Regression (SVR)* didapatkan hasil pengujian pada 3.754 *tweet* menunjukkan *SVR* memperoleh akurasi lebih baik daripada *NB* sebesar 79.7%. Hasil pengujian juga dibandingkan dengan *K-Means Clustering* dan *SVM*. *SVM* memperoleh akurasi sebesar 78.8%, di mana *SVM* lebih baik dari *NB* tetapi masih dibawah *SVR* [25]. Penelitian sebelumnya membuat sistem untuk mendeteksi apakah seseorang terindikasi depresi dan kecemasan atau tidak berdasarkan *Stress Scale - 42 (DASS-42)* dan menggunakan metode *Classification and Regression Tree (CART)* dengan ekstraksi fitur *TF-IDF*. Hasil menunjukkan bahwa model yang paling optimal mencapai nilai akurasi 81.25% dan nilai *F1 Score* 85.71% lebih tinggi dari hasil *baseline* dengan nilai akurasi 62.50% dan nilai *F1 score* 66.66% [26].

Berdasarkan permasalahan yang telah dipaparkan dan beberapa penelitian yang telah dijelaskan, penelitian ini mengusulkan metode *feature extraction* untuk mendeteksi depresi dan kecemasan pada pengguna *Twitter*. Pada penelitian ini juga dilakukan perbandingan dengan beberapa metode *machine learning* sebagai pembanding. Metode pembanding yang digunakan yaitu algoritma *Decision Tree*, *Random Forest*, *Naive Bayes*, dan *K-Nearest Neighbor*. Hasil penelitian ini diharapkan dapat digunakan oleh pembaca maupun pengguna *Twitter* untuk membantu mengidentifikasi penderita depresi dan kecemasan berdasarkan *tweet* dan untuk mengetahui performa algoritma terbaik dalam mengidentifikasi penderita depresi.

2. METODOLOGI PENELITIAN

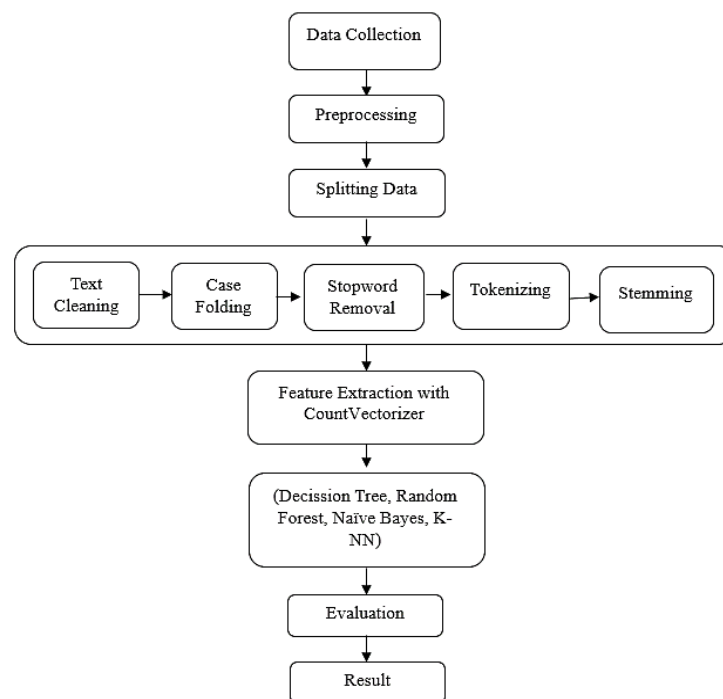
Pada penelitian ini dilakukan perbandingan algoritma klasifikasi teks menggunakan *machine learning* untuk mengidentifikasi penderita depresi dan kecemasan berdasarkan postingan *tweet* pengguna pada *Twitter*. Tahapan penelitian yang dilakukan disajikan pada gambar 1.

2.1 Data Collection

Dataset yang digunakan adalah data sekunder yang diambil dari dataset *Kaggle*, tentang depresi dan kecemasan (link). Dataset terdiri dari 5 atribut (i.e., Text, Label, Age, Gender, Age Category) dan 7489 *tweet* yang diekstraksi menggunakan *Twitter API*. Tabel 1 menyajikan 5 atribut dari dataset yang digunakan dimana pada atribut label 0 menyatakan negatif sedangkan 1 menyatakan positif (0 = negatif, 1 = positif).

2.2 Preprocessing

Proses *preprocessing* ini berfungsi agar data yang akan diproses menjadi lebih terstruktur dan memperlancar jalannya pemodelan. Proses *preprocessing* yang dilakukan pada penelitian ini terdiri dari lima langkah yaitu *text cleaning*, *case folding*, *tokenizing*, *filtering*, dan *stemming*. *Text cleaning* menghilangkan angka, pemisah kata seperti koma (,), titik (.), dan tanda baca lainnya. Tujuan dari *text cleaning* adalah untuk mengurangi noise [22]. *Case folding* adalah langkah *preprocessing* yang tujuannya adalah untuk mengubah semua teks dalam dokumen ke format standar (huruf kecil) [23]. *Tokenizing* memotong *string* input. Beberapa karakter (misalnya tanda baca) dihilangkan dan spasi digunakan sebagai pemisah untuk membagi kalimat menjadi kelompok kata [24]. *Stopword removal* menghilangkan kata-kata yang tidak relevan atau tidak berarti. *Stemming* adalah pencarian kata dasar (*stem word*) yang dihasilkan dari proses *stopword removal* [23].



Gambar 1. Tahapan Penelitian

2.3 Feature Extraction

Perhitungan vektor yang dilakukan pada penelitian ini menggunakan metode *CountVectorizer*. *CountVectorizer* adalah fitur kelas perhitungan numerik dan metode ekstraksi fitur teks yang umum digunakan. *CountVectorizer* mengubah teks kata-kata menjadi matriks frekuensi kata, fungsi matriks *fit_transform* digunakan untuk menghitung jumlah kemunculan setiap kata [21]. Tujuannya adalah untuk memperoleh kata perkata dari semua kalimat dan membangun kosakata dari kata-kata unik yang terdapat pada kalimat tersebut. Kosakata ini digunakan untuk membangun vektor fitur dari banyaknya kata [11].

2.4 Splitting Data

Setelah tahap *preprocessing*, langkah berikutnya adalah memisahkan dataset menjadi data uji dan data latih. Menggunakan pemisahan data 80:20, dataset yang belum pernah digunakan dalam suatu penelitian, tetapi juga berguna untuk mengevaluasi keberhasilan atau kegagalan suatu penelitian, disebut data pengujian, sedangkan data pelatihan yaitu data yang digunakan untuk melakukan penelitian [18].

2.5 Machine Learning

Machine learning atau pembelajaran mesin yaitu salah satu bagian dari kecerdasan buatan yang tujuannya adalah menggunakan mesin dengan cara yang memungkinkan mesin melakukan pekerjaannya dengan cerdas. Karena algoritma pembelajaran mesin memerlukan pengindeksan data, mereka harus terhubung ke database [15]. Model pembelajaran mesin berkaitan erat dengan statistik komputer, dengan tujuan utamanya adalah menggunakan komputer untuk membuat prediksi [16].

2.5.1 Decision Tree

Metode pohon keputusan pembelajaran mesin membuat keputusan pada tingkat yang berbeda menggunakan struktur data pohon ini cocok untuk masalah peramalan karena mudah diinterpretasikan dan

strukturnya stabil. Ini mencakup keduanya klasifikasi (model pohon dengan tujuan yang berbeda untuk kumpulan nilai yang berbeda) dan regresi (tujuan yang mudah berubah). Dalam hal ini, pertanyaan dibagi menjadi dua cabang (ya dan tidak) yang keluar dari pohon. Jumlah yang lebih besar dapat diperoleh pilihan 2 [12].

2.5.2 Random Forest

Random forest adalah metode klasifikasi yang dibuat dengan mengembangkan metode pohon. *Random Forest (RF)* merupakan metode pemisahan biner rekursif untuk mencapai simpul terakhir dari struktur pohon berdasarkan klasifikasi dan pohon regresi [14]. Keputusan didasarkan pada pemilihan atribut secara acak pada setiap node untuk menentukan klasifikasi. Tahap klasifikasi sebagian besar didasarkan pada suara dari pohon keputusan itu sendiri [20].

2.5.3 Naive Bayes

Metode klasifikasi ini berdasar pada *Teorema Bayes* dan mengasumsikan independensi prediktor. Sederhananya, pengklasifikasi *Naive Bayes* berspekulasi bahwa ketersediaan fitur tertentu di kelas tidak ada hubungannya dengan fitur lainnya. *Naive Bayes* terutama berfokus pada industri klasifikasi teks. Ini terutama digunakan untuk mengelompokkan dan mengklasifikasikan objek menurut probabilitas kemunculannya [13].

2.5.4 K-Nearest Neighbor

Algoritma *K-Nearest Neighbor (KNN)* adalah algoritma pembelajaran mesin sederhana yang sering diaplikasikan untuk menyelesaikan masalah pada regresi dan klasifikasi. Klasifikasi *K-Nearest Neighbor (KNN)* merupakan algoritma klasifikasi paling sederhana dan termudah untuk mengelola kumpulan data [19]. Mudah diimplementasikan dan dipahami, tetapi kelemahan utamanya adalah sangat lambat karena ukuran akses data meningkat [13].

2.6 Model Evaluasi

Untuk mengevaluasi kinerja metode *Decision Tree*, *Random Forest*, *Naive Bayes*, dan *K-Nearest Neighbor*, model ini harus diuji. Hasil pengujian disajikan dalam bentuk tabel *confusion matrix*. Skor akurasi model diperoleh dengan membagi jumlah data yang benar pada hasil klasifikasi dengan jumlah total data seperti yang ditunjukkan pada persamaan di bawah ini. Persamaan 1-4 di bawah ini digunakan untuk menghitung tingkat akurasi, recall, presisi dan spesifitas masing masing *confusion matrix*s.

$$Akurasi = (TP + TN) / (TP + FP + FN + TN) \quad (1)$$

$$Recall = (TP) / (TP + FN) \quad (2)$$

$$Precision = (TP) / (TP + FP) \quad (3)$$

$$F1\ Score = 2 * (Recall * Precision) / (Recall + Precision) \quad (4)$$

Keterangan:

- TP (True Positif) = Diagonal matriks
- TN (True Negatif) = Jumlah baris yang sesuai untuk kelas (tidak termasuk TP kelas itu)
- FP (False Positif) = Jumlah kolom yang sesuai untuk kelas (tidak termasuk TP kelas itu)
- FN (False Negatif) = Jumlah dari semua baris dan kolom (tidak termasuk baris dan kolom dari kelas tersebut).

3. HASIL DAN ANALISIS

Pada penelitian ini dataset yang diperlukan adalah dataset publik yang diperoleh dari *website Kaggle*. Dataset terdiri dari 7489 *tweet* di jejaring sosial *Twitter* tentang depresi dan kecemasan. Sebelum menganalisis data, dilakukan langkah preprocessing data, yang meliputi menghapus data *null* atau kosong, menghapus data duplikat, ubah semua teks dalam dokumen menjadi *font* yang konsisten, ubah semua kalimat dalam dokumen sebagai unit kata, hapus kata, *URL*, atau simbol yang tidak bermakna, hapus kata dengan awalan dan akhiran dan mengubah data untuk memenuhi kebutuhan algoritma. Setelah dilakukan *preprocessing* data dan melakukan *feature extraction* didapatkan data terbaru sebanyak 7476 data. Hasil *preprocessing* dapat dilihat pada tabel 1.

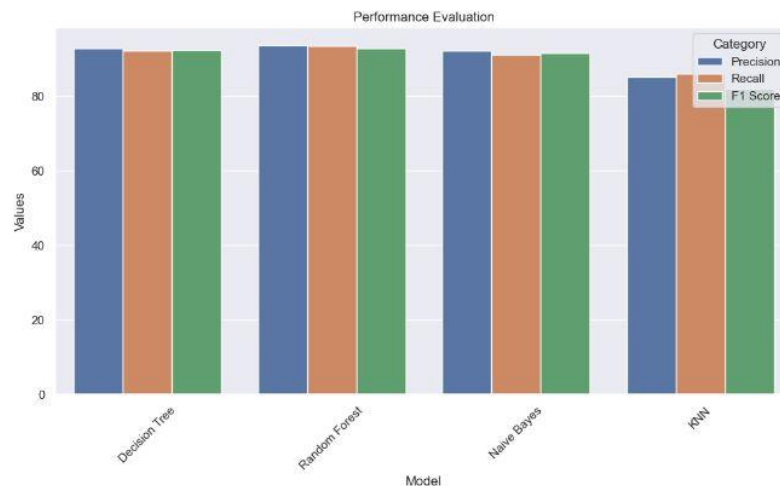
Tabel 1. Hasil Preprocessing

No	Text	Text Cleaning	Case Folding	Stopword Removal	Tokenizing	Stemming
0	let Gether For Party	let Gether For Party	let gether for party	gether party	'gether', 'party'	'gether', 'party'
1	I hate being alive when I feel so dead inside.	I hate being alive when I feel so dead inside	i hate being alive when i feel so dead inside	hate alive feel dead	'hate', 'alive', 'feel', 'dead'	'hate', 'alive', 'feel', 'dead'
...
7474	How come my friends immediately get to know strangers	How come my friends immediately get to know strangers	how come my friends immediately get to know strangers	friends strangers	'friends', 'strangers'	'friends', 'strangers'
7475	Today I wake at 5am.	today I wake at 5am	today i wake at 5am	toady wake 5am	'today', 'wake', '5am'	'today', 'wake', '5am'

Selanjutnya, peneliti mengimplementasikan algoritma *Decision Tree*, *Random Forest*, *Naive Bayes*, dan *K-Nearest Neighbor* dan mengujinya berdasarkan pada tahap pemisahan data 80% dari total data yakni 5980 sebagai data training dan 20% yakni 1496 sebagai data testing berdasarkan parameter teks hasil preprocessing. Hasil komparasi dari keempat algoritma (i.e., *Decision Tree*, *Random Forest*, *Naive Bayes*, dan *K-Nearest Neighbor*) dari penelitian ini pada rasio pemisahan data 80:20 disajikan dalam tabel 2 dan gambar 2.

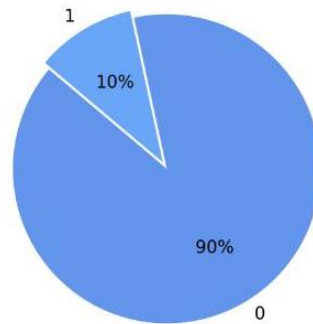
Tabel 2. Komparasi Hasil Algoritma

Algoritma	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.951	92.68	95.86	96.81
Random Forest	0.957	91.93	91.84	90.93
Naive Bayes	0.910	93.45	92.91	93.11
K-Nearest Neighbor	0.858	86.92	86.83	83.42

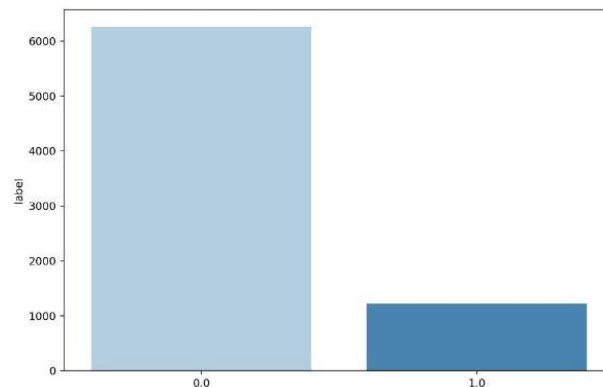
**Gambar 2.** Grafik hasil komparasi algoritma

Pada tabel 2 dan gambar 2 menunjukkan bahwa algoritma yang memiliki tingkat akurasi tertinggi adalah algoritma *Random Forest* dengan tingkat akurasi sebesar 0.957, nilai *precision* 91.93, nilai *recall* 91.84, dan nilai *F1 score* 90.93 dengan menggunakan nilai splitting data sebesar 80:20. Selanjutnya data yang sudah melalui beberapa tahapan akan divisualisasikan, gambar 3 dan gambar 4 adalah hasil visualisasi keseluruhan data yang sudah diproses dan divisualisasikan dengan Diagram *Pie* dan Diagram Batang.

Gambar 4 dapat dilihat bahwa postingan pengguna *Twitter* yang terdeteksi Depresi dan Kecemasan sebanyak 10% yang terdeteksi Positif (1) dan 90% yang terdeteksi Negatif (0). Pada gambar 6, *tweet* yang terdeteksi Positif (1) depresi dan kecemasan sebanyak 1224 *tweet* dan 6252 *tweet* yang terdeteksi Negatif (0). Dapat disimpulkan bahwa hanya sebagian kecil pengguna *Twitter* yang terdeteksi depresi dan kecemasan.



Gambar 3. Diagram Pie



Gambar 4. Diagram Batang

4. KESIMPULAN

Setelah dilakukan tahap implementasi, tahap uji, dan evaluasi yang dilakukan pada sebelumnya dengan menggunakan empat algoritma, yaitu *Decision Tree*, *Random Forest*, *Naive Bayes*, dan *K-Nearest Neighbor (KNN)* didapatkan kesimpulan bahwa tingkat akurasi terbaik terdapat pada algoritma *Random Forest* menggunakan splitting data 80:20 dengan tingkat akurasi sebesar 0.957. Algoritma tersebut dapat secara akurat mengenali pola dan gejala depresi dalam penerapan teknik pemrosesan bahasa alami dan algoritma *machine learning*. Bagi peneliti selanjutnya diharapkan dapat mengembangkan algoritma tersebut dengan menambahkan fitur lainnya, dan disarankan peneliti selanjutnya perlu melakukan tahap penyeimbangan data agar hasil klasifikasi dapat lebih maksimal dengan menggunakan teknik *oversampling* atau *undersampling*.

REFERENSI

- [1] A. Rizki and Y. Sibaroni, "Analisis Sentimen Terhadap Universitas Gunadarma Berdasarkan Opini Pengguna twitter menggunakan metode naive Bayes classifier," *Jurnal Ilmiah Komputasi*, vol. 19, no. 4, 2021.
- [2] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.
- [3] U. Khaira, R. Johanda, P. E. Utomo, and T. Suratno, "Sentiment Analysis of cyberbullying on Twitter using sentiment strength," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 3, no. 1, pp. 21–27, 2020.
- [4] J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T. S. Chua, W. Zhu, Guangyao Shen, "Depression Detection via Harvesting Media : A Multimodal Dictionary Learning Solution," in *Twenty-Sixth International Joint Conference on Artificial Intelligence*.
- [5] D. Septhya, K. Rahayu, S. Rabbani, V. Fitria, R. Rahmaddeni, Y. Irawan, and R. Hayami, Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru: Implementation of Decision Tree Algorithm and Support Vector Machine for Lung Cancer Classification. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol.3, no. 1, pp. 15-19, 2023.
- [6] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah. *E-Bisnis: Jurnal Ilmiah Ekonomi Dan Bisnis*, vol. 13, no. 2, pp. 67-75, 2020.

- [7] C. Ratanamahatana, & D. Gunopulos, Selective Bayesian classifier: feature selection for the Naive Bayesian classifier using decision trees. *WIT Transactions on Information and Communication Technologies*, 2002.
- [8] S. F. Khorshid, and A. M. Abdulazeez, Breast cancer diagnosis based on k-nearest neighbors: a review. *J Arch.Egyptol*, vol. 18, no. 4, pp. 1927-1951, 2021.
- [9] A. Go, R. Bhayani, and L. Huang, Twitter sentiment classification using distant supervision. CS224N project report, Stanford, vol. 1, no. 12, pp. 2009.
- [10] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in Reddit Social Media Forum," *IEEE Access*, vol. 7, pp. 44883–44893, 2019.
- [11] F. P. Rachman, and H. Santoso, Perbandingan Model Deep Learning untuk Klasifikasi Sentiment Analysis dengan Teknik Natural Language Processing. *Jurnal Teknologi dan Manajemen Informatika*, vol. 7, no. 2, pp. 103-112, 2021.
- [12] A. Priya, S. Garg, and N. P. Tigga, "Predicting anxiety, depression and stress in modern life using machine learning algorithms," *Procedia Computer Science*, vol. 167, pp. 1258–1267, 2020.
- [13] B. Mahesh, Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, no. 1, pp. 381-386, 2020.
- [14] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.
- [15] M. Mohammed, M. B. Khan, and B. E. B. Mohammed, *Machine Learning: Algorithms and Applications*. Boca Raton: CRC Press, Taylor & Francis Group, 2016.
- [16] J. Alzubi, A. Nayyar, and A. Kumar, "Machine learning from theory to algorithms: An overview," *Journal of Physics: Conference Series*, vol. 1142, p. 012012, 2018.
- [17] J. L. Ward and J. Peppard, *Strategic Planning for Information Systems*. Chichester: Wiley, 2007.
- [18] M. Nurkholifah, and Y. Umar, ANALISA PERFORMA ALGORITMA MACHINE LEARNING DALAM PREDIKSI PENYAKIT LIVER. *Jurnal Indonesia: Manajemen Informatika dan Komunikasi*, vol. 4, no. 1, pp. 164-172, 2023.
- [19] C. Anuradha, and T. Velmurugan, A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance. *Indian Journal of Science and Technology*, vol. 8, no. 15, pp. 1-12, 2015.
- [20] S. Agustiani, Y. Tajul Arifin, A. Junaidi, S. Khotimatul Wildah, and A. Mustopa, "Klasifikasi Penyakit Daun padi Menggunakan random forest dan color histogram," *Jurnal Komputasi*, vol. 10, no. 1, 2022.
- [21] J. Yang, C. Zhao, H. Yu, and H. Chen, Use GBDT to predict the stock market, *Procedia Computer Science*, vol. 174, pp. 161–171, 2020.
- [22] S. Samsir, A. Ambiyar, U. Verawardina, F. Edi, and R. Watrianthos, "Analisis Sentimen Pembelajaran daring Pada Twitter di masa pandemi COVID-19 Menggunakan metode naïve Bayes," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 1, p. 157, 2021.
- [23] W. A. Luqyana, I. Cholissodin, and R. S. Perdana, Analisis Sentimen Cyberbullying Pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 11, pp. 4704-4713, 2018.
- [24] I. B. Manuaba, G. R. Dantes, and G. Indrawan, "Analisis sentimen data provider Layanan Internet Pada Twitter menggunakan support vector machine Dengan Penambahan Algoritma Levenshtein distance," *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan)*, vol. 5, no. 2, pp. 9–17, 2022.
- [25] P. Arora and P. Arora, "Mining twitter data for Depression Detection," *2019 International Conference on Signal Processing and Communication (ICSC)*, 2019.
- [26] M. R. Hidayatullah and Warih Maharani, "Depression detection on twitter social media using decision tree," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 4, pp. 677–683, 2022.