

Analisis Sentimen Kemungkinan Depresi dan Kecemasan pada Twitter Menggunakan *Support Vector Machine*

Ferry Darmawan¹, Michael Joe², Yogiek Indra Kurniawan*³, Lasmedi Afuan⁴

^{1,2,3,4}Informatika, Fakultas Teknik

Universitas Jenderal Soedirman

Purbalingga, Indonesia

e-mail: ¹ferry.darmawan@mhs.unsoed.ac.id, ²michael.joe@mhs.unsoed.ac.id, ³yogiek@unsoed.ac.id,

⁴lasmedi.afuan@unsoed.ac.id

Diajukan: 24 Januari 2023; Direvisi: 20 Maret 2023; Diterima: 24 Agustus 2023

Abstrak

Lambatnya penanganan terhadap gangguan mental dan adanya pemberian stigma negatif kepada individu yang memiliki gangguan mental menyebabkan mereka memilih untuk menceritakan permasalahannya pada media sosial seperti Twitter. Penelitian ini bertujuan untuk membuat model analisis sentimen yang digunakan untuk melakukan pendeteksian dini pada tweet dari pengguna Twitter yang memiliki tanda gangguan mental berupa depresi atau kecemasan agar dapat segera dilakukan penanganan dengan tepat. Model analisis sentimen yang dibuat menggunakan metode *Support Vector Machine* (SVM). Proses pembuatan model analisis sentimen dilakukan melalui tiga tahapan utama, yaitu pra-pemrosesan data, implementasi SVM, dan analisis hasil. Data yang digunakan adalah data tweet dalam bahasa Indonesia beserta label klasifikasinya. Berdasarkan analisis hasil yang telah dilakukan, hasil klasifikasi terbaik oleh SVM diperoleh dengan menggunakan RBF kernel pada nilai parameter C (complexity) = 10 dan nilai parameter γ (gamma) = 0,1, memiliki nilai akurasi sebesar 82,5%, F1 score sebesar 0,692, nilai precision sebesar 0,715, nilai recall sebesar 0,671, dan rata-rata F1 score sebesar 0,637 pada uji 5-fold cross validation.

Kata kunci: Analisis sentiment, Gangguan mental, Klasifikasi, Support Vector Machine, Twitter.

Abstract

The slow handling of mental disorders and the existence of negative stigma given to individuals who have mental disorders have caused them to choose to share their problems on social media such as Twitter. This study aims to create a sentiment analysis model that is used to perform early detection of tweets from Twitter users who have signs of mental disorders in the form of depression or anxiety so that they can be handled appropriately. Sentiment analysis model created using the *Support Vector Machine* (SVM). The process of making a sentiment analysis model is carried out through three main stages, namely data pre-processing, SVM implementation, and result analysis. The data used is tweet data in Indonesian along with its classification label. Based on the analysis of the results that have been carried out, the best classification results by SVM are obtained using the RBF kernel at the parameter value C (complexity) = 10 and the parameter value (gamma) = 0.1, has an accuracy value of 82.5%, F1 score of 0.692, the precision value is 0.715, the recall value is 0.671, and the average F1 score is 0.637 in the 5-fold cross validation test.

Keywords: Classification, Mental disorders, Sentiment analysis, Support Vector Machine, Twitter.

1. Pendahuluan

Kesehatan mental merupakan suatu kondisi pada individu yang terbebas dari segala bentuk gejala-gejala gangguan mental sehingga individu tersebut dapat memanfaatkan potensi mereka berupa kemampuan kognisi dan emosi untuk berfungsi dalam komunitasnya, memenuhi kebutuhan hidupnya, serta mengatasi tekanan kehidupannya [1], [2]. Kesehatan mental juga merupakan salah satu faktor penting dalam mewujudkan kesehatan individu secara menyeluruh [3]. Maka dari itu, sudah sepantasnya kesehatan

mental mendapatkan perhatian yang sama seperti kesehatan fisik karena kedua hal tersebut saling mempengaruhi [4], [5].

Sama seperti fisik, keadaan mental juga dapat mengalami gangguan kesehatan. Gangguan kesehatan pada mental disebut dengan gangguan mental. Menurut WHO pada tahun 2012, sekitar 350 juta orang lebih mengalami gangguan mental berupa depresi, menjadikan gangguan mental menduduki posisi keempat sebagai penyakit utama di dunia [6]. Tidak seperti gangguan kesehatan pada fisik yang mudah dideteksi karena bersifat tampak, gangguan mental lebih sulit dideteksi karena bersifat tidak tampak. Hal tersebut terkadang menyebabkan lambatnya penanganan terhadap penderita gangguan mental. Selain itu, lambatnya penanganan terhadap penderita gangguan mental juga diperburuk oleh kurangnya kesadaran yang dimiliki oleh masyarakat terkait dengan pentingnya kesehatan mental serta adanya pemberian stigma negatif kepada individu yang memiliki gangguan mental, menyebabkan penderita gangguan mental takut atau malu untuk memeriksakan diri ke fasilitas kesehatan yang ada sehingga memilih untuk menceritakan permasalahannya pada media sosial [7]–[10].

Penggunaan media sosial untuk menceritakan berbagai macam permasalahan terkait dengan gangguan mental yang dialami oleh penderita gangguan mental bukanlah hal yang tepat untuk dilakukan dan tidak akan menyelesaikan masalah yang dialaminya. Penderita gangguan mental harus segera dapat dideteksi sedini mungkin dan ditangani oleh orang yang berkompeten. Penderita gangguan mental yang tidak segera ditangani akan memperparah kondisi kesehatan mental dan bahkan dapat mempengaruhi kondisi kesehatan fisik seperti merasa mudah lelah, jenuh, pusing, mual, dan gangguan pencernaan [2], [4]. Selain itu, gangguan mental dapat menyebabkan penderitanya melakukan tindakan yang berbahaya seperti konsumsi obat terlarang, melukai diri sendiri, atau bahkan bunuh diri [11], [12].

Pada penelitian ini dirancang suatu model analisis sentimen yang digunakan untuk melakukan pendeteksian dini tanda gangguan mental berupa depresi dan kecemasan pada pengguna media sosial Twitter. Pengguna Twitter saling berkomunikasi dan bertukar informasi dalam bentuk *tweet*. Pemilihan Twitter sebagai objek penelitian dikarenakan Twitter adalah salah satu media sosial yang memiliki jumlah pengguna yang banyak karena mudah digunakan sehingga menghasilkan banyak data *tweet* yang dapat dimanfaatkan untuk merancang model analisis sentimen yang optimal [10], [13].

Penelitian terkait analisis sentimen depresi dan kecemasan pada pengguna Twitter pernah dilakukan sebelumnya. Pada penelitian [6], analisis sentimen dilakukan dengan menggunakan metode *deep learning* yang memiliki nilai akurasi sebesar 82.90% dalam performansi klasifikasinya. Pada penelitian [14], analisis sentimen dilakukan dengan menggunakan metode Naïve Bayes Classifier memiliki nilai akurasi sebesar 74% pada pengujian rasio data dan 73% pada pengujian validasi. Pada penelitian [10], dilakukan implementasi Multinomial Naïve Bayes Classifier dengan nilai akurasi sebesar 70% dan nilai *f-measure* sebesar 68% dalam mengklasifikasikan data *tweets* yang mengandung *term* depresi.

Pada penelitian [15], dilakukan analisis sentimen tingkat kepuasan pengguna penyedia layanan telekomunikasi seluler Indonesia pada Twitter menggunakan *SVM* yang memiliki tingkat akurasi sebesar 79% untuk penggunaan Lexicon Based Features dan 84% tanpa penggunaan Lexicon Based Features. Pada penelitian [16], dilakukan analisis sentimen untuk studi tanggapan masyarakat Indonesia terhadap pandemi Covid-19 pada media sosial Twitter. Penggunaan model *SVM* memiliki tingkat akurasi yang lebih baik daripada penggunaan model Naïve Bayes Classifier dan KNN, yaitu 90,01% untuk *SVM* dengan *kernel* linear, 79,20% untuk Naïve Bayes Classifier dengan jumlah *laplace* sebesar 1, dan 62,10% untuk KNN dengan jumlah *K* sebesar 20 dan menggunakan *kernel* optimal. Pada penelitian [17], penggunaan *SVM* untuk analisis sentimen pada Twitter terkait wacana pemindahan ibu kota Indonesia memiliki tingkat akurasi sebesar 96,68%. Pada penelitian [18], penggunaan *SVM* dalam analisis sentimen pada maskapai penerbangan di media sosial Twitter memiliki tingkat akurasi sebesar 84,37% dengan *kernel* RBF, nilai *complexity* sebesar 1, dan *gamma* sebesar 1. Pada penelitian ini, model analisis sentimen yang akan dibuat menggunakan *Support Vector Machine (SVM)*. *SVM* dipilih karena sudah terbukti bekerja cukup baik dalam mengklasifikasikan data berupa teks pada berbagai penelitian yang telah dilakukan sebelumnya.

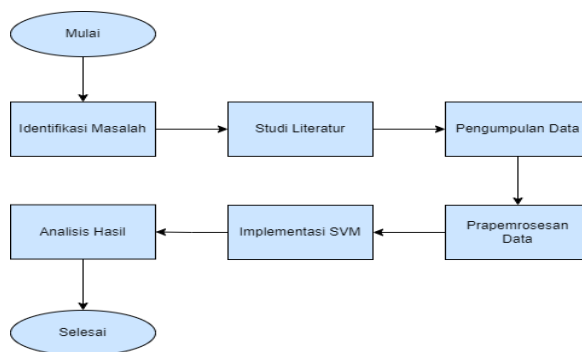
2. Metode Penelitian

Pada penelitian ini digunakan metode penelitian eksperimen. Metode penelitian eksperimen adalah metode penelitian yang digunakan untuk mengamati hubungan kausal antara munculnya suatu akibat dan sebab tertentu dalam kondisi yang terkendali [19]. Langkah-langkah metode eksperimen dalam penelitian ini ditunjukkan pada gambar 1.

2.1. Identifikasi Masalah

Tahap identifikasi masalah dilakukan untuk mengidentifikasi masalah yang ingin diselesaikan dengan dilakukannya suatu penelitian. Masalah yang berhasil diidentifikasi yaitu lambatnya pendeteksian

dan penanganan terhadap penderita gangguan mental yang diperparah dengan kurangnya kesadaran masyarakat akan pentingnya kesehatan mental serta pemberian stigma negatif individu yang memiliki gangguan mental sehingga penderita gangguan mental memilih untuk menceritakan permasalahannya pada media sosial seperti Twitter. Berdasarkan permasalahan tersebut, dibutuhkan suatu solusi yang dapat digunakan untuk melakukan pendeteksian dini gangguan mental agar dapat segera ditangani dengan tepat. Solusi yang dipilih yaitu membuat model analisis sentimen pada *tweet* dari pengguna Twitter yang memiliki tanda gangguan mental berupa depresi atau kecemasan menggunakan *SVM*.



Gambar 1. Metodologi Penelitian

2.2. Studi Literatur

Tahap studi literatur digunakan untuk mencari informasi terkait dengan permasalahan yang diangkat serta solusi yang akan digunakan pada penelitian ini. Informasi tersebut diperoleh dari penelitian-penelitian yang telah dilakukan sebelumnya. Informasi tersebut kemudian dikaji dan dievaluasi agar pada penelitian ini dapat dibuat solusi tepat untuk menyelesaikan permasalahan yang diangkat.

2.3. Pengumpulan Data

Tahap pengumpulan data dilakukan untuk mengumpulkan data yang diperlukan dalam pembuatan model analisis sentimen. Data yang diperlukan adalah *dataset tweet* dalam bahasa Indonesia beserta label klasifikasinya dan *dataset* kamus kata tidak baku (slang) yang digunakan dalam proses *spell normalization*.

Dataset tweet diperoleh dengan cara diunduh melalui internet pada situs penyedia *dataset* bernama Kaggle [20]. *Dataset* yang diunduh memiliki 2701 data *tweet* beserta labelnya. Sampel dari data yang digunakan ditunjukkan pada tabel 1.

Tabel 1. Sampel *Dataset*

No.	Text	Label
1	gelisah bgt astaga	1
2	tidur susah, pikiran kacau, hati gelisah. Semua tak selaras	1
3	Serba salah, mundur sayang, maju ragu. Diem di tempat resah dan gelisah 😞😞😞	1
4	Adek gue sepertinya sudah muak dengerin gue muter sunlit smile	0
5	pagi ini gw wangi banget dah!.wangi minyak telon sama bedak bayi ya Allah dinda 😊	0

Pada tabel 1, nilai 1 pada kolom label berarti pengguna Twitter yang membagikan *tweet* pada kolom *text* menunjukkan tanda potensi kecemasan atau depresi, sedangkan nilai 0 berarti sebaliknya. Distribusi frekuensi nilai label pada *dataset* ditunjukkan pada gambar 2.



Gambar 2. Distribusi Frekuensi Nilai Label pada *Dataset*

Dataset kamus kata tidak baku diperoleh dari penelitian [21]. *Dataset* yang diunduh memiliki 150006 baris kata dan 2 kolom, yaitu kolom kata slang dan kolom kata bakunya. Sampel dari *dataset* kamus kata tidak baku ditunjukkan pada tabel 2.

Tabel 2. Sampel *Dataset* Kamus Kata Tidak Baku

No.	Kata Slang	Kata Baku
1	woww	wow
2	aminn	amin
3	gtau	enggak tau
4	met	selamat

2.4. Prapemrosesan Data

Tahap prapemrosesan data dilakukan untuk meningkatkan kualitas dari data yang sebelumnya telah dikumpulkan sehingga diharapkan hasil dari pembuatan model analisis sentimen menggunakan data tersebut menjadi lebih baik. Teknik prapemrosesan data yang digunakan secara berurutan adalah pembersihan teks, *case folding*, *tokenizing*, *spell normalization*, *filtering*, *stemming*, dan pembobotan kata menggunakan TF-IDF.

2.4.1. Pembersihan Teks

Pembersihan teks dilakukan untuk menghapus kotoran-kotoran pada teks. Kotoran-kotoran pada teks dapat berupa tanda baca, *URL*, *hashtag*, karakter *HTML*, dan emoji. Pada penelitian ini, emoji tidak dihapus karena emoji menyimpan informasi emosi yang tidak dapat disampaikan melalui bahasa tubuh atau karakteristik vokal pada komunikasi nonverbal [22]. Emoji yang ada pada teks diubah ke dalam bentuk ASCII dan diperlakukan seperti teks biasa.

2.4.2. Case Folding

Case folding adalah proses mengubah seluruh huruf pada teks menjadi huruf kecil (lower case) [7], [16]. *Case folding* dilakukan untuk menyamaratakan semua kata pada teks sehingga kata dengan komposisi huruf yang sama dengan perbedaan pada penggunaan jenis hurufnya (huruf kapital dan kecil) akan diberlakukan sama pada tahap prapemrosesan selanjutnya.

2.4.3. Tokenizing

Tokenizing adalah proses memecah kalimat menjadi bagian yang lebih kecil berupa kata-kata penyusunnya [18]. *Tokenizing* dilakukan untuk mempermudah dalam pendeteksian kata yang harus dihilangkan pada proses *filtering* dan pendeteksian kata yang harus diubah pada proses *stemming*. *Tokenizing* juga mempermudah penghitungan frekuensi kemunculan kata pada proses transformasi data ke dalam vektor dokumen.

2.4.4. Spell Normalization

Spell normalization adalah pengubahan kata-kata yang tidak baku ke dalam bentuk kata baku. Kata-kata tidak baku dapat berupa singkatan atau kata-kata slang. *Spell normalization* bertujuan untuk mengurangi keberagaman kata sehingga meningkatkan akurasi [23].

2.4.5. Filtering

Filtering adalah proses menghapus kata-kata yang tidak memiliki makna dan tidak berpengaruh pada proses klasifikasi sentimen berdasarkan daftar *stoplist* yang telah dibuat [6], [17]. *Filtering* dilakukan untuk meningkatkan efisiensi dalam proses pelatihan model analisis sentimen yang dibuat dan dapat meningkatkan akurasi analisis sentimen [24]. Daftar kata yang digunakan pada *stoplist* diambil dari modul bahasa pemrograman Python bernama Sastrawi. Sampel dari *stoplist* ditunjukkan pada tabel 3.

Tabel 3. Sampel *Stoplist*

No.	Kata
1	jika
2	untuk
3	pada
4	ke
5	para

2.4.6. Stemming

Stemming adalah proses pengubahan kata pada kalimat menjadi bentuk dasarnya dengan menghilangkan imbuhan pada kata tersebut [14], [16]. Stemming dilakukan untuk menghilangkan ambiguitas dalam kata [14]. Stemming juga dapat meningkatkan akurasi analisis sentimen dengan cara mengurangi keberagaman kata [25].

2.4.7. Term Frequency – Inverse Document Frequency (TF-IDF)

Seluruh *tweet* yang telah dilakukan pembersihan teks hingga stemming kemudian diubah ke dalam bentuk vektor dokumen. Vektor dokumen adalah sebuah vektor yang berisi kumpulan dokumen dengan atributnya berupa kata-kata yang pernah muncul pada seluruh dokumen tersebut. Pemberian nilai atau bobot pada setiap atribut untuk masing-masing dokumen menggunakan algoritma TF-IDF.

Algoritma TF-IDF (term frequency – inverse document frequency) adalah hasil perkalian dari TF (term frequency) dan IDF (inverse document frequency), dihitung menggunakan persamaan 1.

$$tf - idf(t, d) = tf(t, d) \times idf(t) \tag{1}$$

TF adalah total kemunculan suatu kata dalam suatu dokumen, dengan menggunakan persamaan 2.

$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{2}$$

$f_{t,d}$ adalah jumlah kata t pada dokumen d , sedangkan $f_{t',d}$ adalah jumlah kata t' pada dokumen d . IDF merupakan hasil *inverse* dari perbandingan total dokumen yang memiliki suatu kata dengan total seluruh dokumen, dihitung menggunakan persamaan 3.

$$idf(t) = \log \frac{1 + n}{1 + df(t)} + 1 \tag{3}$$

n pada persamaan 3 merupakan jumlah dokumen dalam *dataset*, sedangkan $df(t)$ adalah jumlah dokumen dalam *dataset* yang mengandung kata t . Hasil dari perhitungan TF-IDF berupa vektor v kemudian dinormalisasi menggunakan persamaan 4.

$$v_{norm} = \frac{v}{\|v\|_2} = \frac{v}{\sqrt{v_1^2 + v_2^2 + \dots + v_n^2}} \tag{4}$$

2.5. Implementasi Support Vector Machine

Support Vector Machine (SVM) adalah suatu teknik pembelajaran berjenis *supervised learning* yang digunakan untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi [15]. *SVM* bekerja dengan cara mencari *hyperplane* atau garis pembatas terbaik sebagai pemisah dua kelas data [15], [18]. *Hyperplane* terbaik adalah *hyperplane* yang memiliki nilai margin maksimal. Margin adalah jarak *support vector* dari masing-masing kelas data ke *hyperplane*. *Support vector* adalah beberapa data terpilih yang digunakan dalam perhitungan untuk membentuk *hyperplane*. *Hyperplane* yang optimal pada linear *kernel* direpresentasikan dengan persamaan (5).

$$(w \cdot x_i) + b = 0 \tag{5}$$

Di dalam data x_i , yang termasuk kelas -1 dapat diperoleh melalui persamaan (6).

$$(w \cdot x_i + b) \leq 1, y_i = -1 \tag{6}$$

Di dalam data x_i , yang termasuk kelas +1 dapat diperoleh melalui persamaan (7).

$$(w \cdot x_i + b) \geq 1, y_i = 1 \tag{7}$$

Dalam menyelesaikan masalah klasifikasi, sering dijumpai kondisi di mana suatu data tidak dapat diklasifikasikan dengan baik secara linear atau data tersebut memiliki dimensi yang banyak. Permasalahan tersebut dapat diselesaikan dengan menggunakan *kernel trick* [18]. Pemilihan *kernel* yang tepat untuk menyelesaikan suatu kasus sangat penting karena mempengaruhi akurasi *SVM* dalam melakukan prediksi. Macam-macam *kernel* dan persamaannya ditunjukkan pada tabel 4.

Tabel 4. Macam-Macam *Kernel* dan Persamaannya

<i>Kernel</i>	Persamaan
<i>Linear</i>	$K(x_i, x) = x_i^T x$
<i>Polynomial</i>	$K(x_i, x) = (\gamma \cdot x_i^T x + r)^p, \gamma > 0$
RBF	$K(x_i, x) = \exp(-\gamma x_i^T - x ^2), \gamma > 0$
<i>Sigmoid</i>	$K(x_i, x) = \tanh(\gamma \cdot x_i^T x + r)$

SVM memiliki *hyperparameter* yang dapat diatur sebelum dilakukan proses latih dan uji. *Hyperparameter SVM* yang digunakan pada penelitian ini adalah *C* (*complexity*), γ (*gamma*) pada non-linear *kernel*, dan *d* (*degree*) pada *polynomial kernel*. *C* adalah parameter yang digunakan untuk mengendalikan berapa banyak data yang sengaja dibiarkan salah terklasifikasi dalam proses latih. γ adalah parameter yang digunakan untuk mengendalikan besarnya pengaruh yang dimiliki oleh suatu data latih yang mempengaruhi kelengkungan pada *hyperplane* yang dibuat. *d* adalah parameter yang digunakan untuk menentukan derajat pada *polynomial kernel* yang mempengaruhi kelengkungan pada *hyperplane* yang dibuat.

2.6. Analisis Hasil

Tahap analisis hasil dilakukan untuk menentukan kualitas dari program analisis sentimen yang dibuat. Analisis dilakukan terhadap nilai akurasi, *F1 score*, *recall*, dan *precision* dari penggunaan *kernel* dengan parameter yang berbeda. Tahap analisis hasil menggunakan beberapa alat bantu seperti *confusion matrix* dan *K-fold cross validation*.

Confusion matrix adalah suatu alat yang digunakan untuk mengevaluasi performa dari model yang telah dibuat dengan membandingkan kelas hasil prediksi dengan kelas sesungguhnya dari suatu data [18]. Evaluasi performa menggunakan *confusion matrix* ditunjukkan pada tabel 5.

Tabel 5. *Confusion Matrix*

Data Aktual	Data Prediksi		
	True	False	Total
True	TP	FN	P
False	FP	TN	N
Total	P'	N'	P + N

TP (*True Positive*) adalah data positif yang terklasifikasi dengan benar, TN (*True Negative*) merupakan data negatif yang terklasifikasi dengan benar, sedangkan FP (*False Positive*) adalah data negatif yang terklasifikasi dengan salah (menjadi positif) dan FN (*False Negative*) adalah data positif yang terklasifikasi dengan salah (menjadi negatif).

Berdasarkan tabel 5, dapat diperoleh persamaan untuk menghitung nilai akurasi, *F1 score*, *recall*, dan *precision* dari suatu model.

Akurasi adalah perbandingan antara data yang terklasifikasi dengan benar dengan jumlah seluruh data, dapat dihitung dengan persamaan 8.

$$Accuracy = \frac{TP + TN}{P + N} \tag{8}$$

F1 score adalah rata-rata harmonik dari *precision* dan *recall*, dapat dihitung dengan persamaan 9.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{9}$$

Precision adalah perbandingan antara data positif yang terklasifikasi dengan benar dengan jumlah seluruh data yang terklasifikasi positif, dapat dihitung dengan persamaan 10.

$$Precision = \frac{TP}{P'} \tag{10}$$

Recall adalah perbandingan antara data positif yang terklasifikasi dengan benar dengan jumlah data dengan label asli positif, dapat dihitung dengan persamaan 11.

$$Recall = \frac{TP}{P} \tag{11}$$

K-fold cross validation yaitu salah satu jenis *cross validation* yang digunakan untuk mengetahui performa model berupa rata-rata keberhasilan dari model yang telah dibuat sehingga model tersebut teruji untuk beberapa data masukan yang berbeda [16], [17]. *K-fold cross validation* bekerja dengan membagi data menjadi sejumlah K kelompok dan melakukan iterasi/perulangan proses latih dan uji sebanyak K dengan data uji adalah kelompok data ke-K dan data latih adalah seluruh data selain data uji. Ilustrasi dari *K-fold cross validation* ditunjukkan pada tabel 6. Sub kolom berwarna gelap pada kolom *dataset* digunakan sebagai data uji, sisanya digunakan sebagai data latih.

Tabel 6. Ilustrasi *K-fold Cross Validation* dengan K = 5

Pengujian	Dataset
1	██████████
2	██████████
3	██████████
4	██████████
5	██████████

3. Hasil dan Pembahasan

Hasil dari penelitian ini adalah sebuah model analisis sentimen menggunakan *SVM* beserta dengan hasil analisis yang dapat memperlihatkan nilai akurasi, *F1 score*, *precision* dan *recall* dalam menyelesaikan permasalahan yang ada. Berikut merupakan hasil dan pembahasan dari penelitian yang telah dilakukan.

3.1. Prapemrosesan Data

Hasil dari proses pembersihan teks sampai *stemming* ditunjukkan pada tabel 7 menggunakan satu sampel dokumen atau kalimat.

Tabel 7. Hasil Proses Pembersihan Teks sampai *Stemming*

No.	Tahap Prapemrosesan	Dokumen
1	-	Ini gelisah bngttt dri kemarin mikirkan bntr lagi kls 12, persiapan sbm masi sedikitt, bnyk bngt yg di khawatirkan 🤔
2	Pembersihan Teks	Ini gelisah bngttt dri kemarin mikirkan bntr lagi kls persiapan sbm masi sedikitt bnyk bngt yg di khawatirkan [DIZZY-FACE]
3	Case folding	ini gelisah bngtt dri kemarin mikirkan bntr lagi kls persiapan sbm masi sedikitt bnyk bgt yg di khawatirkan [DIZZY-FACE]
4	Tokenizing	['ini', 'gelisah', 'bgtt', 'dri', 'kemarin', 'mikirkan', 'bntr', 'lagi', 'kls', 'persiapan', 'sbm', 'masi', 'sedikitt', 'bnyk', 'bngt', 'yg', 'di', 'khawatirkan', '[DIZZY-FACE]']
5	Spell Normalization	['ini', 'gelisah', 'banget', 'dari', 'kemarin', 'mikirkan', 'bentar', 'lagi', 'kls', 'persiapan', 'sbm', 'masih', 'sedikitt', 'banyak', 'banget', 'yang', 'di', 'khawatirkan', '[DIZZY-FACE]']
6	Filtering	['gelisah', 'banget', 'dari', 'kemarin', 'mikirkan', 'bentar', 'kls', 'persiapan', 'sbm', 'masi', 'sedikitt', 'banyak', 'banget', 'yang', 'khawatirkan', '[DIZZY-FACE]']
7	Stemming	['gelisah', 'banget', 'dari', 'kemarin', 'mikirkan', 'bentar', 'kls', 'siap', 'sbm', 'masi', 'sedikitt', 'banyak', 'banget', 'yang', 'khawatir', '[DIZZY-FACE]']

Data yang telah dilakukan pembersihan teks sampai *stemming* kemudian dibagi menjadi data latih dan data uji dengan perbandingan 8:2. Data latih dan data uji yang telah dibuat lalu diubah ke dalam bentuk vektor dokumen, ditunjukkan pada gambar 3.

	0	1	2	3	4	...	5431	5432	5433	5434	5435
0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
...
2746	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
2747	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
2748	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
2749	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
2750	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0

Gambar 3. Vektor Dokumen Data Latih

Pada vektor dokumen, atribut yang berupa seluruh kata yang pernah muncul dalam seluruh dokumen diganti menggunakan angka. Terdapat 5435 atribut yang digunakan, dengan jumlah data latih sebanyak 2201 dan jumlah data uji sebanyak 550.

3.2. Implementasi Support Vector Machine

SVM diimplementasikan menggunakan package *scikit-learn* pada bahasa pemrograman Python [26]. SVM diuji dengan berbagai *kernel* dan *hyperparameter* yang berbeda untuk memperoleh hasil yang terbaik. Penentuan SVM terbaik ditentukan berdasarkan F1 *score* dikarenakan jumlah data yang tidak seimbang antara data *tweet* yang menunjukkan tanda potensi kecemasan atau depresi dengan data yang tidak menunjukkan tanda potensi kecemasan atau depresi.

3.2.1. Linear Kernel

Pada penggunaan linear *kernel*, analisis dilakukan terhadap pengujian parameter C dengan nilai 0,5 sampai 100 dan parameter γ dengan nilai *scale* dan 0,001 sampai 1. Hasil dari pengujian ditunjukkan pada tabel 8.

Berdasarkan tabel 8, penurunan nilai C dari 100 ke 1 memberikan dampak berupa meningkatnya nilai akurasi dan F1 *score*. Namun, penurunan nilai C dari 1 ke 0,5 memberikan dampak yang sebaliknya. Diperoleh model SVM terbaik dari penggunaan *linear kernel* dengan nilai akurasi sebesar 83,3%, F1 *score* sebesar 0,676, nilai *precision* sebesar 0,78, dan nilai *recall* sebesar 0,569. Model tersebut menggunakan nilai parameter C sebesar 1 dan parameter γ sebesar 1.

3.2.2. Polynomial Kernel

Pada penggunaan *linear kernel*, analisis dilakukan terhadap pengujian parameter C dengan nilai 0,5 sampai 100, parameter γ dengan nilai *scale* dan 0,001 sampai 1, dan parameter *d* dengan nilai 2 sampai 4. Hasil dari pengujian ditunjukkan pada tabel 9.

Tabel 8. Hasil Pengujian SVM Menggunakan Linear Kernel

C	γ	Akurasi	F1 score	Precision	Recall	C	γ	Akurasi	F1 score	Precision	Recall
1	0,001	0,782	0,651	0,612	0,696	100	0,1	0,727	0,157	0,824	0,087
1	<i>scale</i>	0,833	0,676	0,78	0,596	100	1	0,727	0,157	0,824	0,087
1	1	0,833	0,676	0,78	0,596	100	<i>scale</i>	0,727	0,157	0,824	0,087
1	0,1	0,778	0,653	0,602	0,714	100	0,001	0,709	0,012	1	0,006
1	0,01	0,833	0,671	0,79	0,584	100	0,01	0,707	0	0	0
10	0,001	0,833	0,671	0,79	0,584	0,5	0,001	0,707	0	0	0
10	0,01	0,736	0,601	0,54	0,677	0,5	0,01	0,707	0	0	0
10	<i>scale</i>	0,735	0,597	0,537	0,671	0,5	0,1	0,707	0	0	0
10	1	0,822	0,611	0,846	0,478	0,5	1	0,707	0	0	0
10	0,1	0,822	0,611	0,846	0,478	0,5	<i>scale</i>	0	0	0	0

Tabel 9. Hasil Pengujian SVM Menggunakan Polynomial Kernel

C	d	γ	Akurasi	F1 Score	Precision	Recall	C	d	γ	Akurasi	F1 Score	Precision	Recall
10	2	1	0,813	0,611	0,779	0,503	10	4	0,1	0,707	0	0	0

10	2	scale	0,813	0,611	0,779	0,503	10	4	0,01	0,707	0	0	0
00	2	1	0,813	0,614	0,774	0,509	10	4	0,001	0,707	0	0	0
00	2	scale	0,813	0,614	0,774	0,509	100	3	0,01	0,707	0	0	0
1	2	scale	0,811	0,567	0,861	0,422	100	3	0,001	0,707	0	0	0
00	2	0,1	0,811	0,567	0,861	0,422	100	2	0,001	0,707	0	0	0
1	2	1	0,811	0,567	0,861	0,422	100	3	0,1	0,707	0	0	0
10	3	scale	0,758	0,393	0,741	0,267	1	4	0,001	0,707	0	0	0
10	3	1	0,758	0,393	0,741	0,267	10	2	0,01	0,707	0	0	0
00	3	1	0,756	0,396	0,721	0,273	0,5	2	0,1	0,707	0	0	0
00	3	scale	0,756	0,396	0,721	0,273	0,5	2	0,01	0,707	0	0	0
0,5	2	scale	0,762	0,342	0,895	0,211	0,5	2	0,001	0,707	0	0	0
0,5	2	1	0,762	0,342	0,895	0,211	0,5	3	0,1	0,707	0	0	0
1	3	scale	0,758	0,345	0,833	0,217	0,5	3	0,01	0,707	0	0	0
1	3	1	0,758	0,345	0,833	0,217	0,5	3	0,001	0,707	0	0	0
10	4	scale	0,738	0,265	0,743	0,161	0,5	4	0,1	0,707	0	0	0
10	4	1	0,738	0,265	0,743	0,161	0,5	4	0,01	0,707	0	0	0
00	4	scale	0,735	0,263	0,703	0,161	0,5	4	0,001	0,707	0	0	0
00	4	1	0,735	0,263	0,703	0,161	1	2	0,1	0,707	0	0	0
1	4	scale	0,74	0,235	0,846	0,137	1	2	0,01	0,707	0	0	0
1	4	1	0,738	0,226	0,84	0,13	1	2	0,001	0,707	0	0	0
0,5	3	1	0,729	0,139	1	0,075	1	3	0,1	0,707	0	0	0
0,5	3	scale	0,731	0,149	1	0,081	1	3	0,01	0,707	0	0	0
0,5	4	scale	0,716	0,06	1	0,031	1	3	0,001	0,707	0	0	0
0,5	4	1	0,716	0,06	1	0,031	1	4	0,1	0,707	0	0	0
00	2	0,01	0,707	0	0	0	1	4	0,01	0,707	0	0	0
10	3	0,1	0,707	0	0	0	100	4	0,01	0,707	0	0	0
10	3	0,01	0,707	0	0	0	10	2	0,1	0,707	0	0	0
10	3	0,001	0,707	0	0	0	10	2	0,001	0,707	0	0	0
00	4	0,1	0,707	0	0	0	100	4	0,001	0,707	0	0	0

Tabel 10. Hasil Pengujian SVM Menggunakan Sigmoid Kernel

C	γ	Akurasi	F1 score	Precision	Recall	C	γ	Akurasi	F1 score	Precision	Recall
10	0,1	0,836	0,688	0,78	0,615	0,5	1	0,815	0,595	0,824	0,466
100	0,01	0,836	0,688	0,78	0,615	10	0,01	0,724	0,136	0,8	0,075
100	0,1	0,784	0,64	0,624	0,658	100	0,001	0,724	0,136	0,8	0,075
10	scale	0,784	0,649	0,618	0,683	1	0,1	0,724	0,136	0,8	0,075
1	scale	0,825	0,655	0,778	0,565	0,5	0,1	0,709	0,012	1	0,006
1	1	0,827	0,659	0,78	0,571	10	0,001	0,707	0	0	0
10	1	0,784	0,649	0,618	0,683	0,5	0,001	0,707	0	0	0
100	1	0,736	0,592	0,541	0,652	0,5	0,01	0,707	0	0	0
100	scale	0,733	0,588	0,536	0,652	1	0,01	0,707	0	0	0
0,5	scale	0,815	0,595	0,824	0,466	1	0,001	0,707	0	0	0

Berdasarkan tabel 9, nilai γ yang terlalu kecil berdampak pada penurunan nilai akurasi dan F1 score. Diperoleh model SVM terbaik dari penggunaan *polynomial kernel* dengan nilai akurasi sebesar 81,3%, F1 score sebesar 0,614, nilai *precision* sebesar 0,774, dan nilai *recall* sebesar 0,509. Model tersebut menggunakan nilai parameter C sebesar 100, nilai parameter *d* sebesar 2, dan nilai parameter γ sebesar 1.

3.2.3. Sigmoid Kernel

Pada penggunaan *Sigmoid kernel*, analisis dilakukan terhadap pengujian parameter C dengan nilai 0,5 sampai 100 dan parameter γ dengan nilai *scale* dan 0,001 sampai 1. Hasil dari pengujian ditunjukkan pada tabel 10.

Berdasarkan tabel 10, nilai γ yang terlalu kecil berdampak pada penurunan nilai akurasi dan F1 score. Diperoleh model SVM terbaik dari penggunaan *sigmoid kernel* dengan nilai akurasi sebesar 83,6%, F1 score sebesar 0,688, nilai *precision* sebesar 0,78, dan nilai *recall* sebesar 0,615. Model tersebut menggunakan nilai parameter C sebesar 10 dan nilai parameter γ sebesar 0,1.

3.2.4. RBF Kernel

Pada penggunaan *RBF kernel*, analisis dilakukan terhadap pengujian parameter C dengan nilai 0,5 sampai 100 dan parameter γ dengan nilai *scale* dan 0,001 sampai 1. Hasil dari pengujian ditunjukkan pada tabel 11.

Tabel 11. Hasil Pengujian SVM Menggunakan RBF Kernel

C	γ	Akurasi	F1 Score	Precision	Recall
10	0,1	0,825	0,692	0,715	0,671
100	0,01	0,82	0,688	0,699	0,677
10	1	0,815	0,638	0,744	0,559
10	<i>scale</i>	0,815	0,638	0,744	0,559
100	1	0,809	0,632	0,726	0,559
100	<i>scale</i>	0,809	0,632	0,726	0,559
100	0,1	0,775	0,622	0,611	0,634
1	1	0,818	0,609	0,821	0,484
1	<i>scale</i>	0,818	0,609	0,821	0,484
100	0,001	0,784	0,441	0,904	0,292

C	γ	Akurasi	F1 Score	Precision	Recall
0,5	1	0,778	0,419	0,898	0,273
0,5	<i>scale</i>	0,778	0,419	0,898	0,273
10	0,01	0,782	0,434	0,902	0,286
1	0,1	0,769	0,38	0,886	0,242
0,5	0,1	0,716	0,082	0,778	0,043
10	0,001	0,707	0	0	0
0,5	0,001	0,707	0	0	0
0,5	0,01	0,707	0	0	0
1	0,01	0,707	0	0	0
1	0,001	0,707	0	0	0

Berdasarkan tabel 11, nilai C dan γ yang terlalu kecil berdampak pada penurunan nilai akurasi dan F1 score. Diperoleh model SVM terbaik dari penggunaan *RBF kernel* dengan nilai akurasi sebesar 82,5%, F1 score sebesar 0,692, nilai *precision* sebesar 0,715, dan nilai *recall* sebesar 0,671. Model tersebut menggunakan nilai parameter C sebesar 10 dan nilai parameter γ sebesar 0,1.

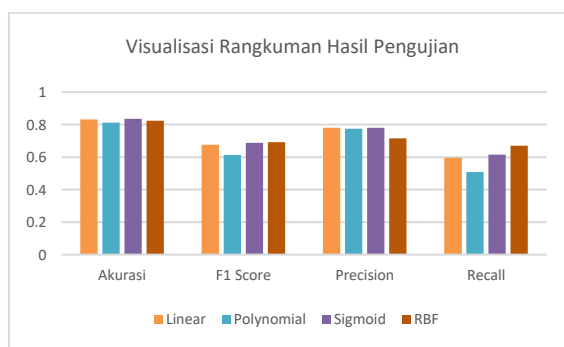
3.3. Analisis Hasil

3.3.1. Pengujian Kernel

Berdasarkan implementasi SVM yang telah dilakukan, diperoleh model terbaik dari masing-masing *kernel*, ditunjukkan pada tabel 12. Visualisasi dari tabel 12 ditunjukkan pada gambar 4.

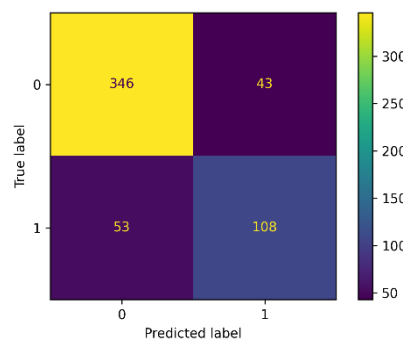
Tabel 12. Rangkuman Hasil Pengujian

Kernel	Parameter	Akurasi	F1 Score	Precision	Recall
Linear	C=1, γ =1	0,833	0,676	0,780	0,596
Polynomial	C=100, d=2, γ =1	0,813	0,614	0,774	0,509
Sigmoid	C=10, γ =0,1	0,836	0,688	0,780	0,615
RBF	C=10, γ =0,1	0,825	0,692	0,715	0,671



Gambar 4. Visualisasi Rangkuman Hasil Pengujian

Berdasarkan tabel 12 dan Gambar 4, penggunaan *RBF kernel* menghasilkan F1 score terbaik, yaitu sebesar 0,692, dengan menggunakan nilai parameter C sebesar 10 dan nilai γ sebesar 0,1. Selain itu, penggunaan *RBF kernel* juga menghasilkan nilai *recall* terbaik, yaitu sebesar 0,671. Namun, penggunaan *RBF kernel* tidak menghasilkan akurasi dan *recall* terbaik. *RBF kernel* memiliki nilai akurasi sebesar 82,5%, menduduki peringkat ketiga di bawah penggunaan *linear kernel* dan *sigmoid kernel*, serta memiliki nilai *precision* sebesar 0,715, terburuk dari seluruh *kernel*. *Confusion matrix* dari model SVM dengan *RBF kernel* terbaik ditunjukkan pada gambar 5.



Gambar 5. Confusion Matrix dari Model SVM dengan RBF Kernel Terbaik

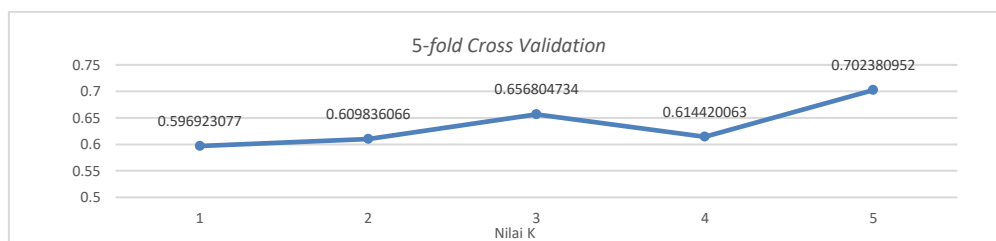
Sampel keluaran dari model SVM dengan RBF kernel terbaik ditunjukkan pada tabel 13.

Tabel 13. Sampel Keluaran Model SVM

No.	Text	Label	Prediksi
1	Ngga enak bgt akhir2 ini rasanya, sering cemas berlebihan	1	1
2	Sedikit khawatir dan gelisah. 😞	1	1
3	Ngelakuin kesalahan yang sama. Bodoh bgt. Udah sebulan ini ngerasa cemas bgt	1	1
4	Pantesan gelisah bae, lupa lgi ga ada duit 😞	0	0
5	Semoga Mei disertai kabar-kabar baik. Yang luk...	0	0

3.3.2. Pengujian K-fold Cross Validation

Pengujian *K-fold cross validation* dilakukan pada model SVM menggunakan RBF kernel dengan nilai parameter C sebesar 10 dan nilai parameter γ sebesar 0,1 untuk memvalidasi F1 score. Nilai K diatur menjadi 5. Hasil dari *5-fold cross validation* ditunjukkan pada gambar 6.



Gambar 6. Hasil 5-fold Cross Validation

Berdasarkan gambar 6, diperoleh rata-rata F1 score sebesar 0,637. Nilai F1-score cenderung naik dengan peningkatan nilai k, dengan nilai F1 score tertinggi didapat sebesar 0.702380952 pada k = 5.

4. Kesimpulan

Berdasarkan hasil dari penelitian yang telah dilakukan maka didapatkan kesimpulan sebagai berikut:

1. SVM terbukti dapat diterapkan dengan baik untuk melakukan analisis sentimen kemungkinan depresi dan kecemasan pada Twitter didukung dengan beberapa metode prapemrosesan dan penentuan nilai *hyperparameter* yang sesuai.
2. F1 score terbaik dihasilkan oleh model SVM dengan penggunaan RBF kernel. Model SVM tersebut memiliki nilai akurasi sebesar 82,5%, F1 score sebesar 0,692, nilai *precision* sebesar 0,715, dan nilai *recall* sebesar 0,671, menggunakan nilai parameter C sebesar 10 dan nilai parameter γ sebesar 0,1.

3. Pengujian *5-fold cross validation* untuk memvalidasi *F1 score* dari model *SVM* yang menggunakan *RBF kernel* dengan nilai parameter *C* sebesar 10 dan nilai parameter γ sebesar 0,1 menghasilkan rata-rata *F1 score* sebesar 0,637.

Daftar Pustaka

- [1] A. W. Putri, B. Wibhawa, and A. S. Gutama, "Kesehatan Mental Masyarakat Indonesia (Pengetahuan, Dan Keterbukaan Masyarakat Terhadap Gangguan Kesehatan Mental)," *Pros. Penelit. dan Pengabd. Kpd. Masy.*, vol. 2, no. 2, pp. 252–258, 2015, doi: 10.24198/jppm.v2i2.13535.
- [2] Y. A. Rozali, N. W. Sitasari, and A. Lenggogeni, "Meningkatkan Kesehatan Mental Di Masa Pandemic," *J. Pengabd. Masy. AbdiMas*, vol. 7, no. 2, 2021, doi: 10.47007/abd.v7i2.3958.
- [3] D. Ayuningtyas, M. Misnaniarti, and M. Rayhani, "Analisis Situasi KesehAyuningtyas, D., Misnaniarti, M., & Rayhani, M. (2018). Analisis Situasi Kesehatan Mental Pada Masyarakat Di Indonesia Dan Strategi Penanggulangannya. *Jurnal Ilmu Kesehatan Masyarakat*, 9(1), 1–10. <https://doi.org/10.26553/jikm.2018.9>," *J. Ilmu Kesehat. Masy.*, vol. 9, no. 1, pp. 1–10, 2018.
- [4] Desti Azania, "Peran Spirtual Bagi Kesehatan Mental Mahasiswa di Tengah Pandemi Covid-19," *Humanistika J. Keislam.*, vol. 7, no. 1, pp. 1–20, 2021.
- [5] M. A. Azhari, "Dukungan Sosial bagi Penderita Disfungsional untuk Penguatan Kesehatan Mental: Studi Syarah Hadis dengan Pendekatan Psikologi Islam," *J. Ris. Agama*, vol. 1, no. 2, pp. 308–322, 2021, doi: 10.15575/jra.v1i2.14569.
- [6] A. Rizki and Y. Sibaroni, "Analisis Sentimen Untuk Pengukuran Tingkat Depresi Pengguna Twitter Menggunakan Deep Learning," *e-Proceeding Eng.*, vol. 8, no. 5, pp. 11367–11375, 2021.
- [7] A. Pragholapati, R. Rizky, A. Shaleha, and L. A. Nasution, "Reduksi Stigma terhadap Orang Dengan Gangguan Mental," *J. Paradig.*, vol. 3, no. 2, pp. 1–6, 2021.
- [8] W. W. Ariesty, Y. E. Praptiningsih, and M. Kasfi, "Sistem Pakar Diagnosa Kesehatan Mental," *JIKI (Jurnal Ilmu Komput. Informatika)*, vol. 2, no. 1, pp. 80–89, 2021, doi: 10.24127/jiki.v2i1.1096.
- [9] D. M. Putri, Yasir, and Nurjanah, "Etnografi Virtual Pengungkapan Diri Mengenai Self Harm Dan Kesehatan Mental Pada Pengguna Media Sosial Twitter," *PERSEPSI Commun. J.*, vol. 4, no. 2, pp. 221–235, 2021.
- [10] N. Hayatin, "Implementasi Multinomial Naïve Bayes Untuk Klasifikasi Data *Tweets* Mengandung Term Depresi," *Pros. SENTRA (Seminar Teknol. dan Rekayasa)*, vol. 0, no. 6, pp. 344–349, 2021.
- [11] Nagara, Widiyanti, Hidayat, and Kurniawan, "Emosional Freedom Technique untuk Kesehatan Mental," *Q. J. Heal. Psychol.*, vol. 8, no. 32, pp. 73–92, 2020.
- [12] D. Esterlina *et al.*, "Perubahan Mental Dengan Memanfaatkan Startup," vol. 2, no. 1, pp. 147–151, 2022.
- [13] K. Yan, D. Arisandi, P. Studi, S. Informasi, and U. Tarumanagara, "ANALISIS SENTIMEN KOMENTAR NETIZEN TWITTER TERHADAP KESEHATAN MENTAL MASYARAKAT INDONESIA," *Junral Ilmu Komput. dan Sist. Inf.*, vol. 10, no. 1, pp. 1–8, 2022, doi: <http://dx.doi.org/10.24912/jiksi.v10i1.17865>.
- [14] M. M. Nurrochman and A. L. Prasasti, "Implementasi Machine Learning Untuk Mendeteksi Unsur Depresi Pada *Tweet* Menggunakan Metode Naïve Bayes (Machine Learning Implementation for Depression Detection in *Tweet* Using Naïve Bayes Method)," *e-Proceeding Eng.*, vol. 8, no. 5, pp. 6250–6257, 2021.
- [15] R. S. Perdana and M. A. Fauzi, "Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter dengan Metode *Support Vector Machine* dan Lexicon Based Features Twitter event detection View project Human Detection

- and Tracking View project,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1(12), no. October, pp. 1725–1732, 2017.
- [16] F. Sodik and I. Kharisudin, “Analisis Sentimen dengan SVM , NAIVE BAYES dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter,” *Prisma*, vol. 4, pp. 628–634, 2021.
- [17] 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.
- [18] H. C. Husada and A. S. Paramita, “Analisis Sentimen Pada Maskapai Penerbangan di Platform Twitter Menggunakan Algoritma *Support Vector Machine (SVM)*,” *Teknika*, vol. 10, no. 1, pp. 18–26, 2021, doi: 10.34148/teknika.v10i1.311.
- [19] A. Jaedun, “Metode Penelitian Eksperimen,” *Metodol. Penelit. Eksperimen*, no. 1, pp. 0–12, 2022.
- [20] S. Hans, “Depression and Anxiety in Twitter (ID),” 2021. .
- [21] N. Aliyah Salsabila, Y. Ardhito Winatmoko, A. Akbar Septiandri, and A. Jamal, “Colloquial Indonesian Lexicon,” *Proc. 2018 Int. Conf. Asian Lang. Process. IALP 2018*, pp. 226–229, 2019, doi: 10.1109/IALP.2018.8629151.
- [22] M. Mostafavi and M. D. Porter, “How emoji and word embedding helps to unveil emotional transitions during online messaging,” *15th Annu. IEEE Int. Syst. Conf. SysCon 2021 - Proc.*, 2021, doi: 10.1109/SysCon48628.2021.9447137.
- [23] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, “An unsupervised lexical normalization for Roman Hindi and Urdu sentiment analysis,” *Inf. Process. Manag.*, vol. 57, no. 6, p. 102368, 2020, doi: 10.1016/j.ipm.2020.102368.
- [24] A. Santosa, I. Purnamasari, and R. Mayasari, “Pengaruh Stopword Removal dan Stemming Terhadap Performa Klasifikasi Teks Komentar Kebijakan New Normal Menggunakan Algoritma,” *Junral Sains Komput.*, vol. 6, no. 1, pp. 81–93, 2022, doi: <http://dx.doi.org/10.30645/j-sakti.v6i1.427>.
- [25] Y. A. Alhaj, J. Xiang, D. Zhao, M. A. A. Al-Qaness, M. Abd Elaziz, and A. Dahou, “A Study of the Effects of Stemming Strategies on Arabic Document Classification,” *IEEE Access*, vol. 7, pp. 32664–32671, 2019, doi: 10.1109/ACCESS.2019.2903331.
- [26] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.