MATHunesa
Jurnal Ilmiah Matematika ***Volume 10 No 01***
e-ISSN : 2716-506X | p-ISSN : 2301-9115 ***Tahun 2022***

KLASIFIKASI EMOSI PADA CUITAN DI TWITTER DENGAN PRINCIPAL COMPONENT ANALYSIS DAN SUPPORT VECTOR MACHINE

Abi Nizar Sutranggono
Jurusan Matematika, FMIPA, Universitas Negeri Surabaya
e-mail: abi.18052@mhs.unesa.ac.id

Abstrak

Salah satu platform media sosial dengan total pengguna aktif harian terbesar adalah Twitter. Melalui Twitter, orang-orang bisa membagikan suatu pesan yang disebut dengan tweet. Ungkapan yang diekspresikan pada tweet dapat merefleksikan bagaimana emosi atau perasaan yang dimiliki seseorang. Emosi yang terkandung dalam sebuah tweet bisa dikenali lewat proses analisis sentimen. Namun, data teks Twitter tidak terstruktur, mengingat saat ini penggunaan singkatan kata, emoji, atau bahkan frasa khusus banyak dijumpai pada tweet, termasuk tweet yang diunggah oleh masyarakat Indonesia. Sehingga, untuk mengidentifikasi emosi dari data teks Twitter melalui proses analisis sentimen dibutuhkan penerapan metode yang tepat. Di sisi lain, *Machine Learning* telah banyak diaplikasikan dalam melakukan tugas analisis sentimen. Kerangka kerja yang disajikan pada penelitian ini melibatkan penggunaan dari algoritma *Machine Learning* untuk dapat menganalisis emosi yang dimuat tweet berbahasa indonesia. Selebihnya, implementasi metode *FastText* dan teknik ekstraksi fitur PCA juga diterapkan agar *output* yang diberikan maksimal. Secara keseluruhan hasil penelitian menunjukkan bahwa *classifier Support Vector Machine* (SVM) dengan fungsi kernel RBF yang dikombinasikan menggunakan PCA memiliki kinerja yang unggul dalam mengklasifikasikan emosi pada tweet berbahasa indonesia, dimana berturut-turut *Accuracy*, *Precision*, *Recall*, serta *F1 Score* yang dicapai sebesar 70,52%, 74,60%, 69,80%, dan juga 71,20%.
Kata Kunci: Klasifikasi Emosi, Tweet Indonesia, Principal Component Analysis, Support Vector Machine.

Abstract

Twitter is one of the social media platforms with the most total daily active users. A tweet is a message that can be shared on Twitter. Expressions channeled through tweets reflect how someone is dealing with emotions or feelings. A sentiment analysis process can be used to identify the emotions contained in a tweet. Twitter text data is not structured, given the prevalence of abbreviations, emojis, and even special phrases in tweets, including those posted by Indonesians. To identify emotions from Twitter text data using the sentiment analysis process, the correct method must be used. Machine Learning, on the other hand, has been widely used in sentiment analysis tasks. The framework presented in this study makes use of Machine Learning algorithms to analyze emotions in Indonesian-language tweets. In addition, the FastText method and the PCA feature extraction technique are used to ensure that the maximum output is obtained. Overall, the research results show that the Support Vector Machine (SVM) classifier with the RBF kernel function combined with PCA performs better in classifying emotions in Indonesian-language tweets, with Accuracy, Precision, Recall, and F1 Score of 70.52%, 74.60%, 69.80%, and 71.20%, respectively.
Keywords: *Emotion Classification, Indonesian Tweets, Principal Component Analysis, Support Vector Machine.*

PENDAHULUAN

Jejaring sosial telah mengalami perkembangan yang pesat. Ada berbagai informasi yang disebarkan oleh pengguna media sosial. Informasi tersebut dapat berupa video, foto, audio ataupun format teks. Di Indonesia, media sosial yang biasanya dipergunakan oleh masyarakat umum antara lain Twitter, Facebook dan Instagram. Selain itu, salah satu platform media sosial dengan pengguna aktif harian terbanyak yaitu Twitter. Fitur yang disediakan Twitter memudahkan pengguna untuk berbagi pesan yang disebut dengan cuitan atau tweet. Isi dari setiap tweet juga sangatlah beragam. Ada tweet yang berisikan keluhan, opini, bahkan kritikan negatif dan positif. Konten tweet ini bervariasi, sehingga ilmuwan, pemerintah, dan suatu perusahaan mencoba memproses informasi Twitter untuk mendapatkan wawasan. Selain itu, platform media sosial seperti Twitter bisa digunakan sebagai media untuk mendapatkan suatu data yang dapat mendukung penelitian di bidang psikologi dan ilmu perilaku. (Murphy, 2017). Tweet yang diunggah bisa mencerminkan bagaimana emosi atau perasaan yang dimiliki oleh seseorang. Maka dari itu, informasi dari Twitter juga bisa dipakai untuk memahami kondisi mental dari individu (Lin, 2015).

Pada dasarnya, lewat ekspresi wajah serta ucapan, emosi manusia dapat diterjemahkan secara langsung. Umumnya, emosi seseorang dikelompokkan menjadi dua kategori, positif dan negatif. Emosi negatif terdiri dari *sadness*, *fear*, *anger*, serta *disgust*. Pada sisi lainnya, emosi positif yakni *happy* atau *joy* (Izard, 2009). Emosi yang termuat di dalam suatu postingan media sosial bisa dikenali melalui sebuah proses analisis sentimen. Analisis sentimen melibatkan pengaplikasian metode *Natural Language Processing* (NLP) untuk menafsirkan emosi dari seseorang atau kelompok secara sistematis (Medhat dkk., 2014).

Pemanfaatan *machine learning* dalam suatu analisis sentimen dengan menggunakan data Twitter masih menjadi tren suatu penelitian (Drus & Khalid, 2019). Kasus yang diangkat serta algoritma yang digunakan beragam. Salah satu studi melakukan proses analisis sentimen dengan tujuan untuk memahami pendapat orang-orang mengenai kegiatan belajar secara *online* selama pandemi dengan menggunakan algoritma *K-Nearest Neighbors*, *Random Forest*, dan juga algoritma *Naïve Bayes* (Althagafi dkk., 2021). *Machine learning* dapat diterapkan untuk menjalankan deteksi emosi dan klasifikasi sentimen dari suatu tweet (Sailunaz & Alhajj, 2019). Karenanya, kegunaan *machine Learning* bisa juga diterapkan untuk melakukan analisis emosi, khususnya guna mempelajari suatu kondisi psikologi seseorang berdasarkan cuitan (Nandwani & Verma, 2021).

Untuk melakukan analisis emosi yang termuat pada sebuah data teks memakai *machine learning*, maka perlu diaplikasikan teknik *Feature Engineering* (Carrillo-de-Albornoz dkk., 2018). Salah satu metode *Feature Engineering* yang biasanya digunakan dalam tugas analisis sentimen yaitu *word embedding*. Metode *Word2Vec Embedding* dapat diimplementasikan untuk merepresentasikan kata pada teks ke dalam bentuk numerik (Mikolov dkk., 2013). Namun, metode ini mempunyai keterbatasan, metode *Word2Vec* tidak mampu memetakkan kata ke dalam suatu vektor dengan baik jika suatu kata tidak termuat dalam sebuah *corpus* (Bojanowski dkk., 2017). Sementara itu, metode *FastText Embedding* bisa mengatasi isu *Out of Vocabulary* (OOV) ini (Kwon dkk., 2021).

Feature Extraction adalah sebuah proses yang tidak kalah penting dalam melakukan klasifikasi memakai algoritma *Machine Learning*. Teknik *Feature Extraction* digunakan untuk mendapatkan hasil klasifikasi yang lebih baik. Sebuah studi mengkombinasikan metode ekstraksi fitur *Principal Component Analysis* (PCA) dan algoritma *Support Vector Machine* untuk menganalisis data teks Twitter (Anjaria & Guddeti, 2014). Pada sisi lain, hasil studi menunjukkan bahwa penerapan dari PCA dapat membantu mereduksi dimensi fitur yang digunakan serta meningkatkan akurasi klasifikasi. Di samping itu, metode PCA juga dapat mempersingkat waktu pemrosesan analisis sentimen (Cheng & Chen, 2019). PCA adalah teknik ekstraksi fitur yang populer dan efektif digunakan di berbagai bidang (Vinodhini & Chandrasekaran, 2014).

Mengidentifikasi emosi berdasarkan data Twitter adalah sebuah pekerjaan yang menantang (Kharde & Sonawane, 2016). Hal tersebut dikarenakan data teks Twitter tidak terstruktur. Penggunaan dari singkatan kata, emoji, atau bahkan frasa khusus saat ini banyak dijumpai dalam suatu tweet. Pada Twitter Indonesia, juga ditemui masalah serupa. Tweet yang disebarkan oleh masyarakat mengandung kata informal (Saputri dkk., 2018). Selain itu, menemukan representasi fitur dari data teks yang tepat serta mewakili suatu entitas tidaklah mudah (Cheng & Chen, 2019).

Penelitian ini dilakukan analisis sentimen dengan menerapkan algoritma berbasis *machine learning* guna mengidentifikasi suatu emosi yang terkandung pada tweet berbahasa indonesia. Secara khusus algoritma yang dipergunakan yaitu *K-Nearest Neighbors* (KNN), *Random Forest*, serta *Support Vector Machine* (SVM). Pada sisi lainnya, untuk memperoleh hasil klasifikasi yang optimal, penerapan dari metode *FastText* dan ekstraksi fitur PCA juga digunakan dalam penelitian yang dilakukan.

Sistematika penyusunan artikel ini yaitu: Bagian 2 menyajikan kajian teori yang meliputi penjelasan dari metode *word embedding*, PCA, dan algoritma *Machine Learning*. Dataset dan proses analisis sentimen yang dilakukan pada penelitian ini ditunjukkan di bagian 3. Bagian 4 memaparkan hasil eksperimen. Terakhir, simpulan dan saran dijelaskan pada bagian 5.

KAJIAN TEORI

WORD EMBEDDING

Teknik ekstraksi atau konversi fitur dari data teks menjadi bentuk angka disebut w*ord embedding* (Ruder & Søgaard, 2019). *Word embedding* diterapkan untuk memetakkan suatu kata atau frasa pada suatu teks ke dalam sebuah vektor bilangan *real* dengan dimensi *N* (Chandra & Krishna, 2021). Selain itu, penerapan dari teknik *word embedding* bertujuan agar mempermudah

proses klasifikasi teks (Li & Yang, 2018). Di samping itu, *FastText* adalah salah satu teknik *word embedding*. FastText merupakan bentuk pengembangan metode *Word2Vec* yang mampu mengatasi permasalahan *Out of Vocabulary* (OOV) dengan cara menerapkan konsep *similarity* makna kata (Khattak dkk., 2019). Sehingga, kata dengan makna yang sama dengan kata yang ada pada suatu *corpus* dapat dicari representasi numerik walaupun bentuk katanya berbeda.

PRINCIPAL COMPONENT ANALYSIS (PCA)

Metode yang dapat digunakan untuk mereduksi dimensi dari data input ke dalam dimensi yang lebih kecil dengan mempertimbangkan variansi data yaitu metode PCA. PCA diaplikasikan untuk menemukan bentuk transformasi linier yang memetakkan sebuah kerangka koordinat dari data input ke koordinat baru yang berbentuk ortonormal (Cushion dkk., 2019). Di sisi lainnya, *Principal Components* (PCs) adalah suatu transformasi linier dari data input menjadi data baru Ukuran data yang ingin disimpan direpresentasikan oleh pengambilan dari PCs. Dengan PCA, algoritma ekstraksi fitur berikut ini:

- Data input untuk diekstrak fiturnya disiapkan.
- Data input dinormaliasi.
- Suatu matriks kovarians dihitung.

$$Cov(xy) = \frac{\sum xy}{n} - (\bar{x})(\bar{y}) \quad (1)$$

- Nilai eigen dan juga vektor eigen dihitung.

$$(A - \lambda I) = 0 \quad (2)$$

- Nilai eigen diurutkan dari yang terbesar ke yang terkecil (*descending*) serta menentukan PCs.

$$[A - \lambda I][X] = [0] \quad (3)$$

K-NEAREST NEIGHBORS (KNN)

KNN (*K-Nearest Neighbors*) merupakan algoritma yang paling sederhana dari sekian algoritma *Artificial Learning* (Diz dkk., 2016). Cara kerja dari algoritma ini yaitu dengan mempertimbangkan jarak terdekat dari suatu sampel. Untuk mengklasifikasikan objek, KNN memakai nilai k tetangga terdekat sebagai acuan agar dapat menentukan kelas dari objek tersebut. Pada sisi lainnya, dalam kasus *binary classification*, nilai k yang dipilih berupa bilangan ganjil. Hal tersebut bertujuan untuk menghindari banyaknya kelas yang sama dari tetangga terdekat (Cherif, 2018). Jarak terdekat dapat dicari dengan menghitung jarak *Euclidean*, *City Block*, dan *Chebyshev*.

$$Euclidean\ D(x,p) = \sqrt{(x-p)^2} \quad (4)$$

$$City\ Block\ D(x,p) = |x-p| \quad (5)$$

$$Chebyshev\ D(x,p) = Max(|x-p|) \quad (6)$$

RANDOM FOREST

Algoritma klasifikasi yang diaplikasikan dengan menggunakan metode *bootstrap aggregating* (*bagging*) dan juga *random feature selection* adalah *Random Forest* (Breiman, 2001). Cara kerja dari *Random Forest* adalah dengan menganalisis suatu pohon-pohon (*tress*) yang ditumbuhkan dan menjadi hutan (*forest*). Penentuan klasifikasi dengan memakai algoritma *Random Forest* diperoleh berdasarkan hasil suatu *majority voting* dari setiap pohon (Kulkarni & Sinha, 2013).

Dalam membentuk sebuah hutan yang terdiri dari *random tree*, tahap yang dilibatkan sebagai berikut:

1. Dengan melakukan suatu penggantian terhadap gugus data (*bootstrap*), N kasus dipilih acak.
2. Pohon disusun hingga tumbuh mencapai ukuran maksimal tanpa melewati pemangkasan dengan menggunakan contoh *bootstrap*. Di setiap pohon, atribut dengan jumlah m juga dipilih secara acak.
3. Tahap 1 dan 2 diulangi sampai k kali, sehingga terbentuk hutan yang terdiri dari k pohon.

SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) merupakan salah satu algoritma berbasis *machine learning* yang sering digunakan dalam tugas klasifikasi di berbagai bidang (Demidova dkk., 2016). Di sisi lain, *Non-linear* SVM mengklasifikasikan suatu data dengan menentukan *hyperplane* terbaik yang disusun dari kumpulan data yang menjadi suatu pemisah di antara setiap kelas. Kumpulan data tersebut adalah titik-titik terdekat pada sebuah *hyperplane* yang disebut *support vector* (Imah dkk., 2011). Selain itu, *hyperplane* terbaik dapat dicari dengan menghitung nilai *margin* maksimum, dimana *margin* merepresentasikan jarak dari kelas satu ke kelas lain. Performa SVM dapat dipengaruhi oleh parameter yang telah ditentukan. Selengkapnya, fungsi kernel yang digunakan adalah parameter yang dimaksud. Dalam algoritma SVM terdapat beberapa macam fungsi kernel, contohnya *Linear Kernel*, *Radial Basis Function* (RBF) *Kernel*, serta *Polynomial Kernel*. Secara matematis fungsi kernel tersebut didefinisikan berikut ini:

$$K(x,y) = (x^T y) \quad (7)$$

$$K(x,y) = exp\left\{\frac{\|x-y\|^2}{2\sigma^2}\right\} \quad (8)$$

$$K(x,y) = (x^T y + c)^d \quad (9)$$

persamaan (7), (8), dan juga (9) secara berturut-turut adalah fungsi kernel linier, RBF, serta polinomial.

Secara khusus, x, y adalah input data, c yaitu suatu nilai konstan, d merupakan derajat polinomial, dan σ adalah suatu standar deviasi.

Misal terdapat dataset yang sudah dilabeli dalam dua kelas yakni: $x_i \in R^d$, $y_i \in \{-1, +1\}$ dimana $i = \mathbb{N}$, $d > 1$, dan *hyperplane* $g(x) = \langle w, x \rangle + c$, maka sebuah keputusan ditentukan memakai definisi:

$$g(x) < 1 \longrightarrow y_i = -1 \quad (10)$$

$$g(x) \geq 1 \longrightarrow y_i = +1 \quad (11)$$

Nilai *margin* dimaksimalkan untuk mengetahui *hyperplane* terbaik:

$$\frac{1}{2}\|w\|^2 = \frac{1}{2}(w_1^2 + w_2^2) \quad (12)$$

dengan ketentuan yang berlaku adalah:

$$(w_1 x_i + w_2 x_i + c) \geq 1 \quad (13)$$

Suatu fungsi *sign decision* $f(x)$ dihitung supaya dapat menentukan kelas dari suatu data.

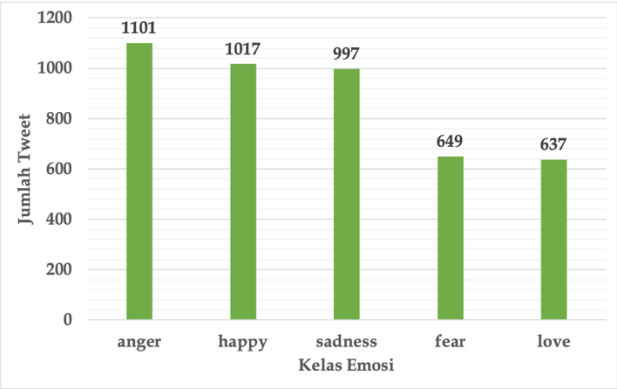
$$f(x_d) = \sum_{i=1}^{m} \alpha_i \, y_i K(x_i, x) + y \quad (14)$$

dimana m adalah *support vectors*, α_i adalah bobot setiap data, dan $K(x_i, x)$ adalah fungsi kernel.

METODE

DATASET

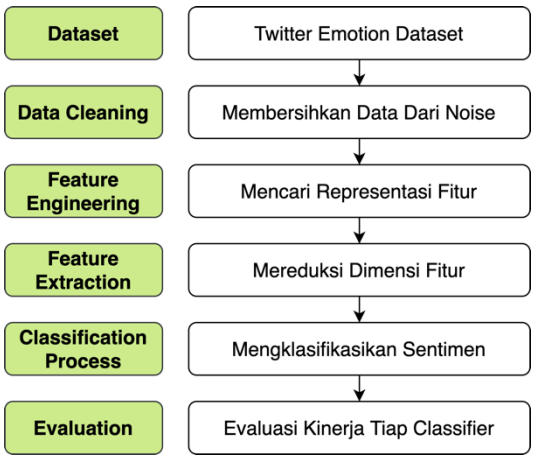
Twitter Emotion Dataset adalah suatu dataset yang dipakai pada penelitian ini (Saputri dkk., 2018). Data tersebut adalah kumpulan tweet berbahasa indonesia yang dikoleksi memakai teknik *Streaming* via Twitter API sejak tanggal 1 Juni 2018 sampai 14 Juni 2018. Jumlah dari keseluruhan data yakni sebanyak 4.401 tweet. Di samping itu, setiap tweet telah dilabeli oleh tim peneliti ke dalam lima kelas emosi yang berbeda, yaitu *anger*, *happy*, *sadness*, *fear*, dan *love*. Berdasarkan Gambar 1, tweet yang terdapat pada dataset tersebut rata-rata dikelompokkan ke dalam kelas emosi *anger*, dimana jumlah tweetnya sebanyak 1101 tweet.



Gambar 1. Distribusi Jumlah Tweet Per Kelas Emosi

ANALISIS SENTIMEN

Untuk melakukan suatu analisis sentimen dengan menggunakan *Machine Learning*, dalam penelitian ini serangkaian proses yang dilibatkan yaitu, antara lain *Data Cleaning*, *Feature Engineering*, *Feature Extraction*, *Classification Process*, dan juga *Evaluations*. Gambar 2, menggambarkan suatu proses analisis sentimen dari penelitian yang dijalankan. Secara khusus, penelitian ini terdiri atas dua eksperimen: 1) tanpa menerapkan proses *Feature Extraction*; serta 2) dengan menerapkan *Feature Extraction*.



Gambar 2. Proses Analisis Sentimen

Langkah pertama dalam analisis sentimen adalah melakukan *Data Cleaning* atau *Preprocessing*. Data teks Twitter yang dipergunakan masih memuat beberapa *noise*. Karenanya, sebelum mengaplikasikan *Machine Learning* data tersebut perlu diproses terlebih dahulu. Proses *Data Cleaning* yang diterapkan meliputi:

- *Removal of Sensitive Information*, yaitu menghapus informasi sensitif seperti *mentions*, *hashtags*, URL, serta *sensitive numbers* sesuai dengan format dari dataset.
- *Emoticons Transformation*, mentransformasi setiap *emoticon* menjadi token string yang bersesuaian.
- *Removal of Punctuations*, yakni menghapus tanda baca yang terdapat dalam suatu tweet.
- *Lower Casing*, mengubah huruf pada setiap tweet menjadi huruf kecil (*lowercase*).
- *Informal Words Transformation*, merupakan proses transformasi kata yang tidak formal menjadi kata baku.
- *Removal of Single Letters*, penghapusan *single letter* atau huruf tunggal yang masih ada pada teks.
- *Removal of Stopwords*, menghapus setiap kosakata yang tidak memberikan informasi penting dalam proses klasifikasi.

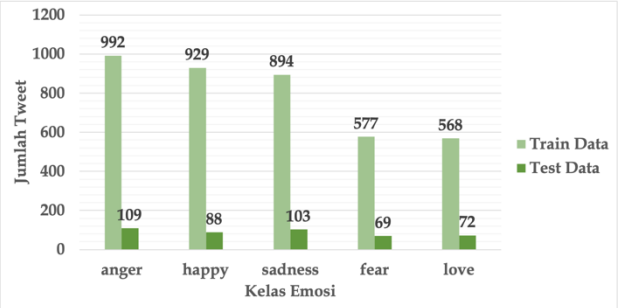
Tabel 1. Contoh Transformasi Kata

Emoticon & Slang Words	Transformasi Kata
:O	terkejut
:)	senyum
:p	mengejek
sans	santai
cemunguth	semangat
eps	episode

Selanjutnya, usai melakukan *Data Cleaning*, proses berikutnya adalah *Feature Engineering* yang bertujuan untuk mengonversi data berupa teks menjadi sebuah vektor bilangan *real* dengan mengaplikasikan metode berbasis *word embedding*, yakni *FastText*. Di samping itu, dalam penelitian ini *FastText* diimplementasikan dengan menggunakan *pre-trained FastText model* yang berdimensi 100 (Saputri dkk., 2018). Pada sisi lainnya, *pre-trained model* tersebut memuat 69.465 vektor kata.

Data yang sudah dikonversi mempunyai dimensi yang cukup besar, yaitu 4401×100. Sehingga, untuk mereduksi dimensi dari fitur yang dipakai dilakukan proses *Feature Extraction*. Metode ekstraksi fitur yang diterapkan yakni *Principal Component Analysis* (PCA). Dengan menggunakan PCA, dimensi dari data input diperkecil menjadi 4401×80.

Kemudian, setelah melewati proses ekstraksi fitur data input dibagi menjadi 90% *train data* dan 10% *test data*, dimana distribusi dari pembagian data tersebut disajikan pada Gambar 3. Selain itu, dalam penelitian ini juga dilakukan normalisasi pada masing-masing data yang diolah. Untuk melakukan klasifikasi emosi dari suatu tweet, algoritma yang diimplementasikan yaitu algoritma KNN, *Random Forest*, dan juga SVM. Di sisi lain, pada tahapan *Classification Process*, setiap model klasifikasi (*classifiers*) yang dirancang, dilatih dengan memakai *train data* untuk pelatihan (*training*).



Gambar 3. Distribusi Data Train dan Test

Ketiga *classifier* tersebut diujikan dengan memakai *test data* untuk melakukan analisis emosi. Kinerja dari masing-masing *classifier* dievaluasi guna mengetahui seberapa besar keakuratan model dalam menganalisa sentimen. Adapun tolak ukur yang digunakan untuk mengevaluasi kinerja dari setiap *classifier* yaitu antara lain hasil akurasi dalam memprediksi data serta lama waktu yang dibutuhkan *classifier* dalam menganalisa sentimen (*testing time*).

METRIK EVALUASI

Evaluation metrics yang digunakan pada penelitian ini yakni, *Accuracy*, *Precision*, *Recall*, dan juga *F1 Score*. Selebihnya, masing-masing *evaluation metrics* tersebut dirumuskan sebagai berikut:

$$Accuray = \frac{TP+TN}{(TP+FN)+(FP+TN)} \quad (15)$$

$$Precision = \frac{TP}{TP+FP} \quad (16)$$

$$Recall = \frac{TP}{TP+FN} \quad (17)$$

$$F1\ Score = \frac{2TP}{2TP+FN+FP} \quad (18)$$

Dengan TP, TN, FP, dan FN secara berturut-turut adalah *True Positive*, *True Negative*, *False Positive*, serta *False Negative*.

HASIL DAN PEMBAHASAN

Penelitian ini mengklasifikasikan tweet berbahasa indonesia ke dalam lima kelas emosi yang mencakup: *anger*, *happy*, *sadness*, *fear*, dan *love*. Untuk melakukan tugas tersebut, algoritma klasifikasi yang digunakan yakni, algoritma KNN, *Random Forest*, dan juga SVM. Di samping itu eksperimen yang dijalankan meliputi: 1) melakukan klasifikasi tanpa menerapkan ekstraksi fitur; dan 2) mengimplementasikan metode ekstraksi fitur PCA. Di sisi lainnya, algoritma *Machine Learning* dipergunakan dengan pengaturan yang disesuaikan. Algoritma KNN diaplikasikan dengan memakai nilai $k = 5$. Sementara itu, beberapa macam fungsi kernel digunakan dalam menerapkan algoritma SVM. Tabel 2, menunjukkan hasil evaluasi dari tiap model dalam melakukan klasifikasi, dimana proses ekstraksi fitur tidak dilibatkan.

Tabel 2. Hasil Evaluasi Klasifikasi Tanpa Ekstraksi Fitur

Classifier	Accuracy	Testing Time
KNN	55,78%	0,08 detik
Random Forest	57,37%	0,09 detik
SVM *Linear*	64,40%	0,07 detik
SVM *RBF*	64,63%	0,34 detik
SVM *Polynomial*	65,99%	0,09 detik

Nilai akurasi tertinggi pada Tabel 2 yakni sebesar 65,99% menggunakan algoritma SVM dengan fungsi kernel polinomial berderajat 2. Selain itu, lama waktu

yang dibutuhkan SVM *Polynomial* dalam melakukan klasifikasi adalah 0,09 detik. Di sisi lain, berdasarkan Tabel 2, nilai akurasi terendah yaitu 55,78% memakai KNN, dan lama waktu pemrosesan terlama diperoleh menggunakan SVM dengan fungsi kernel RBF, yaitu 0,34 detik.

Tabel 3. Hasil Evaluasi Klasifikasi Dengan Ekstraksi Fitur

Classifier	Accuracy	Testing Time
KNN	58,50%	0,07 detik
Random Forest	62,36%	0,08 detik
SVM *Linear*	65,99%	0,06 detik
SVM *RBF*	70,52%	0,34 detik
SVM *Polynomial*	67,80%	0,09 detik

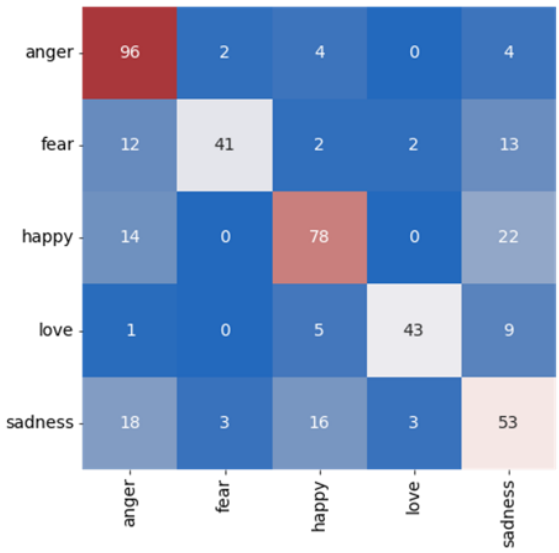
Principal Component Analysis (PCA) diaplikasikan untuk mereduksi dimensi dari fitur yang digunakan. Pada penelitian ini, fitur dengan dimensi 4401 × 100 diperkecil menjadi 4401 × 80. Hasil evaluasi masing-masing model dalam melakukan klasifikasi sentimen setelah menerapkan ekstraksi fitur ditunjukkan pada Tabel 3. Akurasi tertinggi didapatkan menggunakan SVM dengan fungsi kernel RBF, yaitu sebesar 70,52%. Sedangkan lama waktu yang dibutuhkan oleh model tersebut untuk mengeksekusi adalah 0,34 detik. Nilai akurasi terendah dimiliki KNN, yakni 58,50%. Selain itu, lama waktu pemrosesan tercepat diraih oleh SVM dengan fungsi kernel linier, 0,06 detik. Performa dari masing-masing *classifier* secara keseluruhan menjadi lebih optimal sesudah menerapkan PCA. Selebihnya, peneliti mencatat apabila dimensi dari fitur direduksi lagi, maka hasil yang diberikan tidak maksimal. Jadi, dimensi fitur yang proporsional adalah 4401 × 80. Di sisi lainnya, *Principal Components* yang dipakai, yakni 80, memuat 94,24% dari karakteristik data input.

Eksperimen yang sudah dikerjakan menunjukkan bahwa setelah mengimplementasikan ekstraksi fitur, performa *classifier* SVM *RBF* lebih unggul dibanding dengan *classifier* lainnya, walaupun lama waktu yang dibutuhkan untuk mengeksekusi tidak singkat, akan tetapi akurasi yang dicapai cukup baik, yakni sebesar 70,52%. E*valuation metrics* SVM dengan fungsi kernel RBF yang dikombinasikan dengan PCA ditunjukkan Tabel 4. Model SVM dengan fungsi kernel *Radial Basis Function* (RBF) yang dikombinasikan menggunakan PCA mempunyai nilai *Precision* 74,60%, *Recall* 69,80%, serta *F1 Score* 71,20%.

Tabel 4. Evaluation Metrics SVM RBF Dengan PCA

Class	Precision	Recall	F1 Score
anger	68,00%	91,00%	78,00%
happy	74,00%	68,00%	71,00%
sadness	52,00%	57,00%	55,00%
fear	89,00%	59,00%	71,00%
love	90,00%	74,00%	81,00%
Average	74,60%	69,80%	71,20%

Selain meninjau *evaluation metrics* dan *testing time*, untuk mengukur kinerja dari suatu model klasifikasi, penelitian ini juga menganalisis hasil *confusion matrix* atau matriks konfusi *classifier* terbaik. Dari Gambar 4, dapat dipahami bahwa, *classifier* SVM dengan fungsi kernel RBF yang dirancang menggunakan PCA, bisa memprediksi benar kelas *anger* sebanyak 96 dari 109. Kelas *fear* terprediksi benar sebanyak 41 dari 69. Pada sisi lain, kelas *happy* terdapat 78 diprediksi benar dari 88. Kemudian kelas *love* diprediksi benar sebanyak 43 dari 72. Dan kelas *sadness* terprediksi benar sebanyak 53 dari 103.



Gambar 4. Confusion Matrix SVM RBF Dengan PCA

KESIMPULAN

Penelitian ini telah melibatkan penggunaan dari algoritma *Machine Learning* antara lain, KNN, *Random Forest*, serta SVM untuk melakukan klasifikasi emosi yang termuat dalam tweet berbahasa indonesia. Pada sisi lain, berdasarkan hasil eksperimen ditunjukkan dengan mengimplementasikan teknik ekstraksi fitur menggunakan metode *Principal Component Analysis* (PCA), mampu mengoptimalkan suatu peforma dari

masing-masing *classifiers*, sehingga dapat diperoleh hasil klasifikasi yang lebih akurat. Adapun c*lassifier* terbaik adalah SVM dengan memakai fungsi kernel RBF. Akurasi tertinggi yang diperoleh yaitu sebesar 70,52%. Di samping itu, berturut-turut nilai *Precision*, *Recall*, dan juga *F1 Score* dari *classifer* tersebut sebesar 74,60%, 69,80%, serta 71,20%. Selebihnya, penerapan dari PCA juga dapat mempercepat kinerja dari setiap *classifier*.

Peneliti menekankan bahwa proses *Data Cleaning* memainkan peran yang sangat penting dalam proses analisis sentimen, dimana prosedur yang diterapkan harus menyesuaikan karakteristik dataset. Selain itu, peneliti berharap penelitian ini dapat dikembangkan dengan menerapkan metode reduksi dimensi lainnya sehingga bisa diperoleh hasil akurasi yang lebih baik lagi.

DAFTAR PUSTAKA

Althagafi, A., Althobaiti, G., Alhakami, H., & Alsubait, T. (2021). Arabic Tweets Sentiment Analysis about Online Learning during COVID-19 in Saudi Arabia. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 12, Issue 3). https://textblob.readthedocs.io/en/dev/

Anjaria, M., & Guddeti, R. M. R. (2014). Influence Factor Based Opinion Mining of Twitter Data Using Supervised Learning. *Sixth International Conference on Communication Systems and Networks (COMSNETS)*, 1–8.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. http://www.isthe.com/chongo/tech/comp/fnv

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Carrillo-de-Albornoz, J., Vidal, J. R., & Plaza, L. (2018). Feature engineering for sentiment analysis in e-health forums. *PLoS ONE*, *13*(11). https://doi.org/10.1371/journal.pone.0207996

Chandra, R., & Krishna, A. (2021). COVID-19 sentiment analysis via deep learning during the rise of novel cases. *PLoS ONE*, *16*(8 August). https://doi.org/10.1371/journal.pone.0255615

Cheng, C. H., & Chen, H. H. (2019). Sentimental text mining based on an additional features method for text classification. *PLoS ONE*, *14*(6). https://doi.org/10.1371/journal.pone.0217591

Cherif, W. (2018). Optimization of K-NN algorithm by clustering and reliability coefficients: Application to breast-cancer diagnosis. *Procedia Computer Science*, *127*, 293–299. https://doi.org/10.1016/j.procs.2018.01.125

Cushion, E. J., Warmenhoven, J., North, J. S., & Cleather, D. J. (2019). Principal component analysis reveals the proximal to distal pattern in vertical jumping is governed by two functional degrees of freedom. *Frontiers in Bioengineering and Biotechnology*, *7*(AUG). https://doi.org/10.3389/fbioe.2019.00193

Demidova, L., Nikulchev, E., & Sokolova, Y. (2016). Big Data Classification Using the SVM Classifiers with the Modified Particle Swarm Optimization and the SVM Ensembles. *International Journal of Advanced Computer Science and Applications*, *7*(5). https://doi.org/10.14569/ijacsa.2016.070541

Diz, J., Marreiros, G., & Freitas, A. (2016). Applying Data Mining Techniques to Improve Breast Cancer Diagnosis. *Journal of Medical Systems*, *40*(9). https://doi.org/10.1007/s10916-016-0561-y

Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, *161*, 707–714. https://doi.org/10.1016/j.procs.2019.11.174

Imah, E. M., Afif, F. al, Fanany, M. I., & Basaruddin, T. (2011). A Comparative Study on Daubechies Wavelet Transformation, Kernel PCA and PCA as Feature Extractors for Arrhythmia Detection Using SVM. *TENCON 2011 - 2011 IEEE Region 10 Conference*, 5–9. https://doi.org/https://doi.org/10.1109/TENCON.2011.6129052

Izard, C. E. (2009). Emotion theory and research: Highlights, unanswered questions, and emerging issues. In *Annual Review of Psychology* (Vol. 60, pp. 1–25). https://doi.org/10.1146/annurev.psych.60.110707.163539

Kharde, V. A., & Sonawane, S. S. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. In *International Journal of Computer Applications* (Vol. 139, Issue 11). http://ai.stanford.

Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., & Rudzicz, F. (2019). A survey of word embeddings for clinical text. In *Journal of Biomedical Informatics: X* (Vol. 4). Academic Press Inc. https://doi.org/10.1016/j.yjbinx.2019.100057

Kulkarni, V. Y., & Sinha, P. K. (2013). Random Forest Classifiers: A Survey and Future Research Directions. In *International Journal of Advanced Computing* (Vol. 36, Issue 1).

Kwon, O., Kim, D., Lee, S.-R., Choi, J., & Lee, S. (2021). Handling Out-Of-Vocabulary Problem in Hangeul Word Embeddings. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 3213–3221. https://doi.org/10.18653/v1/2021.eacl-main.280

Li, Y., & Yang, T. (2018). *Word Embedding for Understanding Natural Language: A Survey* (pp. 83–104). https://doi.org/10.1007/978-3-319-53817-4_4

Lin, J. (2015). On Building Better Mousetraps and Understanding the Human Condition: Reflections on Big Data in the Social Sciences. *Annals of the*

American Academy of Political and Social Science, *659*(1), 33–47. https://doi.org/10.1177/0002716215569174

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality*. http://arxiv.org/abs/1310.4546

Murphy, S. C. (2017). A Hands-On Guide to Conducting Psychological Research on Twitter. *Social Psychological and Personality Science*, *8*(4), 396–412. https://doi.org/10.1177/1948550617697178

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. In *Social Network Analysis and Mining* (Vol. 11, Issue 1). Springer. https://doi.org/10.1007/s13278-021-00776-6

Ruder, S., & Søgaard, A. (2019). A Survey of Cross-lingual Word Embedding Models. In *Journal of Artificial Intelligence Research* (Vol. 65). http://labs.theguardian.com/digital-language-divide/

Sailunaz, K., & Alhajj, R. (2019). Emotion and sentiment analysis from Twitter text. *Journal of Computational Science*, *36*. https://doi.org/10.1016/j.jocs.2019.05.009

Saputri, M., Mahendra, R., & Adriani, M. (2018). Emotion Classification on Indonesian Twitter Dataset. *International Conference on Asian Language Processing 2018*, 90–95.

Vinodhini, G., & Chandrasekaran, R. M. (2014). Opinion mining using principal component analysis-based ensemble model for e-commerce application. *CSI Transactions on ICT*, *2*(3), 169–179. https://doi.org/10.1007/s40012-014-0055-3