Klasifikasi Emosi dalam Tweet Twitter Berbahasa Indonesia Menggunakan Regresi Logistik dengan Hashing Vectorizer dan Variance Threshold

Fiskal Purbawan[1], Endang Sugiharti[2]

[1,2]Jurusan Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang
Email: [1]purbawanfiskal@students.unnes.ac.id, [2]endangsugiharti@mail.unnes.ac.id

Abstrak

Media sosial saat ini telah menjadi salah satu indikator dalam pengambilan keputusan. Salah satu teknik dalam memanfaatkan data di media sosial saat ini adalah analisis emosi berbasis teks. Analisis emosi berbasis teks banyak digunakan oleh perusahaan dan pemerintah untuk dipertimbangkan dalam pengambilan keputusan dan strategi pemasaran. Tujuan dari penelitian ini adalah untuk menerapkan Hashing Vectorizer sebagai vektorisasi kata dan Variance Threshold sebagai seleksi fitur dalam algoritma Logistic Regression pada tweet-tweet berbahasa Indonesia di Twitter. Dataset yang digunakan adalah Indonesian-Twitter-Emotion-Dataset yang dibagi menjadi dua bagian, yaitu data pelatihan dan data uji dengan proporsi 70:30. Data pelatihan diterapkan dengan teknik Hashing Vectorizer dan Variance Threshold dan kemudian diproses menggunakan algoritma Regresi Logistik. Selain itu, model yang telah dibentuk diuji menggunakan data uji dan kemudian dievaluasi menggunakan laporan klasifikasi dan matriks kebingungan untuk mendapatkan nilai akurasi. Akurasi yang dihasilkan oleh metode ini adalah 62%.
Kata Kunci: Regresi Logistik, Hashing Vectorizer, Ambang Varians, Klasifikasi Emosi, Twitter

Abstract

Social media today has now become one of the indicators in decision-making. One of the techniques in utilizing data on social media today is text-based emotion analysis. Text-based emotion analysis is widely used by companies and governments to be considered in decision-making and marketing strategies. The purpose of this research is to apply Hashing Vectorizer as word vectorization and Variance Threshold as feature selection in Logistic Regression algorithm on Indonesian tweets on Twitter. The dataset used is Indonesian-Twitter-Emotion-Dataset which is divided into two parts, namely training data and test data with a proportion of 70:30. The training data is applied with Hashing Vectorizer and Variance Threshold techniques and then processed using the Logistic Regression algorithm. Furthermore, the model that has been formed is tested using test data and then evaluated using a classification report and confusion matrix to get an accuracy value. The accuracy produced by this method is 62%.
***Keyword**: Logistic Regression, Hashing Vectorizer, Variance Threshold, Emotion Classification, Twitter*

1. PENDAHULUAN

Perkembangan pesat teknologi internet saat ini memberikan pengguna akses mudah untuk pertukaran informasi. Sebagai salah satu media untuk pertukaran informasi, media sosial dapat menyediakan informasi dalam bentuk opini, sentimen, dan emosi

dari penggunanya. Dari pertukaran informasi ini, hal ini kemudian akan tercermin pada bagaimana keselarasan pengguna media sosial terhadap suatu entitas, acara, dan kebijakan [1]. Salah satu media sosial terkenal dengan banyak pengguna adalah Twitter.

Berdasarkan laporan yang dipublikasikan oleh we are social dan hootsuite pada Juli 2021, jumlah pengguna aktif di media sosial twitter per Januari 2021 mencapai 353 juta pengguna. Dalam laporan yang sama, jumlah pengguna twitter di Indonesia mencapai 14,05 juta pengguna. Angka ini menempatkan twitter di posisi kelima sebagai media sosial yang banyak digunakan oleh masyarakat Indonesia [2]. Besarnya jumlah pengguna media sosial khususnya twitter di Indonesia tidak lepas dari kondisi Indonesia sebagai negara kepulauan yang kaya akan keragaman dan kepadatan penduduk yang tinggi [3].

Saat ini, data telah menjadi aset yang penting, termasuk data dari media sosial. Media sosial merupakan salah satu sumber informasi yang memegang peranan penting bagi perusahaan atau pemerintah dalam mengambil keputusan, manajemen administrasi, kampanye politik dan lain-lain. Media sosial dapat memberikan pengaruh yang besar bagi kehidupan masyarakat [4]. Media sosial berbasis microblogging seperti twitter telah menjadi media yang kuat dan telah digunakan oleh jutaan pengguna di jejaring sosial [5]. Konten di twitter, atau yang sering disebut dengan kicauan, telah banyak digunakan oleh para peneliti, pemerintah, dan industri. Penggunaan data dari tweet ini bertujuan untuk mendapatkan pengetahuan yang diharapkan dapat membantu memecahkan masalah sehari-hari. Berbagai aktivitas dan kebiasaan aktual seseorang dapat dilihat dan direkam melalui kicauannya. Salah satu aplikasi yang paling populer adalah analisis emosi [6].

Emosi adalah suatu keadaan pikiran yang berkelanjutan, yang ditandai dengan gejala mental, fisik, dan perilaku [7]. Emosi seseorang dapat diidentifikasi secara langsung melalui ekspresi wajah dan ucapan. Deteksi emosi secara otomatis menjadi penting karena dapat diimplementasikan di berbagai bidang. Dalam bidang pendidikan, misalnya, analisis emosi dapat digunakan dalam lingkungan e-learning yang cerdas [8]. Selain itu, analisis emosi dapat digunakan dalam bidang bisnis untuk mengidentifikasi keluhan pelanggan melalui surat elektronik [9].

Di era sekarang ini, orang dapat mengekspresikan emosinya melalui teks termasuk melalui unggahan di media sosial. Di media sosial seperti Twitter, deteksi emosi dapat berguna bagi pemerintah untuk memantau respon masyarakat terkait dengan pembuatan kebijakan atau kegiatan politik. Selain pemerintah, analisis emosi di media sosial juga digunakan oleh perusahaan untuk memantau bagaimana respon masyarakat terhadap layanan atau produk mereka. Analisis emosi juga digunakan oleh perusahaan untuk menentukan target pasar mereka [6].

Beberapa peneliti telah menjadikan deteksi emosi sebagai topik bahasan, salah satunya adalah masalah deteksi emosi pada tweet berbahasa Indonesia. Penelitian yang dilakukan oleh [6] mengumpulkan data tweet bahasa Indonesia dari twitter kemudian membaginya ke dalam lima kelas, yaitu marah, takut, gembira, cinta, dan sedih. [6] menggunakan beberapa metode yaitu random forest, logistic regression, dan support vector machine dengan menggunakan berbagai macam ekstraksi fitur seperti

lexicon-based, post tagging, ortographic dan kombinasi dari semua metode ekstraksi fitur. Hasil dari penelitian tersebut menunjukkan bahwa hasil terbaik didapatkan dari kombinasi algoritma logistic regression dengan kombinasi semua ekstraksi fitur dengan 75.98% pada dataset lama, dan 69.73% pada dataset baru.

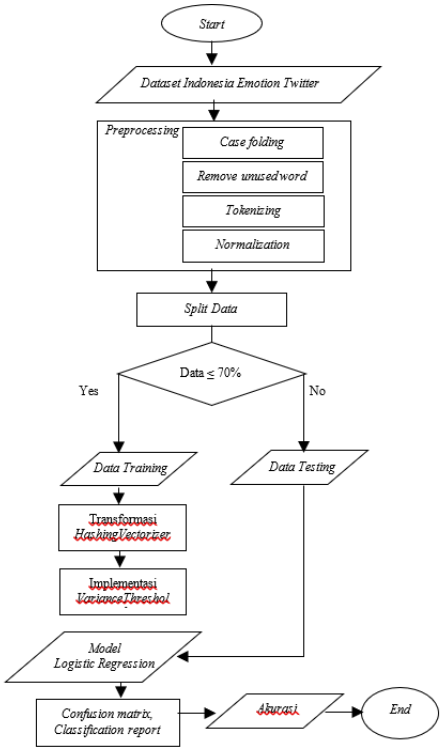
Kemudian penelitian terkait penggunaan berbagai jenis metode vektorisasi data kata untuk analisis sentimen, salah satunya adalah hashing vectorizer. Penelitian yang dilakukan oleh Haque dkk. (2019) yang melakukan opinion mining pada ulasan restoran berbahasa Bangla. Pada penelitian ini digunakan teknik N-gram sebagai ekstraksi fitur dan hashing vectorizer, count vectorizer dan term frequency-inverse document frequency sebagai pembobotan teks. Hasil dari penelitian ini menunjukkan bahwa hashing vectorizer dapat memberikan akurasi yang tinggi jika n fitur juga diatur pada angka yang tinggi [10].

Penelitian tentang ambang batas varians sebagai pemilihan fitur dalam algoritma klasifikasi belum banyak digunakan dalam analisis emosi berbasis teks. Namun, hal ini telah dilakukan pada bidang penelitian lain. Salah satunya adalah dalam bidang computer vision. Salah satunya pada penelitian yang dilakukan oleh Siti Ambarwati dan Uyun (2020). Penelitian tersebut dilakukan dengan menggunakan variance threshold sebagai seleksi fitur pada algoritma k-nearest neighbor (knn) untuk proses pengklasifikasian telur. Hasilnya adalah variance threshold sebagai seleksi fitur dapat meningkatkan akurasi algoritma knn pada sistem egg candling [11].

Berdasarkan deskripsi di atas, penelitian ini mencoba untuk fokus pada analisis penelitian klasifikasi emosi pada data teks tweet dari Twitter. Algoritma Regresi Logistik digunakan sebagai pembelajaran mesin. Vektorisasi kata dari data menggunakan hashing vectorizer dengan ambang varians sebagai pemilihan fitur. Sehingga penelitian ini mengambil judul "Klasifikasi Emosi dalam Teks Tweet Twitter di Indonesia Menggunakan Regresi Logistik dengan Hashing Vectorizer dan Variance Threshold.

2. METODE

Dalam penelitian ini, klasifikasi emosi berdasarkan teks bahasa Indonesia dengan menerapkan Hashing Vectorizer sebagai transformasi data. Kemudian Variance Threshold digunakan sebagai pemilihan fitur. Dan penggunaan regresi logistik untuk melakukan proses klasifikasi. Penilaian tingkat akurasi metode ini menggunakan teknik matriks kebingungan dan laporan klasifikasi. Alur penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. *Flowchart* Alur Penelitian

Penelitian ini dilakukan dalam beberapa tahap. Dimulai dengan proses pra-pemrosesan dataset yang mencakup pembersihan, pengubahan huruf besar-kecil, normalisasi, stemming, penghapusan kata-kata umum, dan tokenisasi. Kemudian dilanjutkan dengan menerapkan hashing vectorizer dan variance threshold.an proses klasifikasi menggunakan regresi logistik.

2.1. Dataset

Data yang digunakan dalam penelitian ini berasal dari penelitian yang dilakukan oleh [6] yang disebut indonesian-twitter-emotion-dataset. Dataset ini terdiri dari satu atribut kelas yang terdiri dari lima kelas yaitu anger, fear, joy, love, and sadness. Contoh data dari dataset yang digunakan dapat dilihat di Tabel 1.

Tabel 1. Sampel dataset

Emotion Labels	Tweet
anger	"Soal jln Jatibaru,polisi tdk bs GERTAK gubernur .Emangny polisi tdk ikut pmbhasan? Jgn berpolitik. Pengaturan wilayah,hak gubernur. Persoalan Tn Abang soal turun temurun.Pelik.Perlu kesabaran. [USERNAME] [USERNAME] [URL]"
sadness	"Orang lain kalau pake ponco itu buat jas hujan, nah dia pake buat kasur. Ya tadi gara2 saking gak punya apa2. Mamak bilang, kami tuh di

	awal pernikan gak ada ngalamin bulan madu kayak skrg2. Org tidur nya aja pake ponco. Gimane mau bulan madu."
...	…
happy	[USERNAME] dulu beneran ada mahasiswa Teknik UI nembak pacarnya pas sahur di Kukusan Teknik Depok, diliput kru Katakan Cinta (dan belum pacaran mereka). Sekarang udah nikah dan punya anak. Pernah diceritain laman UI Shitposting/Divarposting juga.

2.2. Preprocessing

2.2.1 Cleaning

Tahap cleaning adalah di mana tweet dihapus dari karakter yang tidak perlu. Contohnya adalah simbol, angka, dan tanda baca. Dalam dataset ini, selain mengandung simbol, angka, dan tanda baca, data nama pengguna dan tautan juga terdapat dalam tweet. Data username dan tautan ini telah diubah ke dalam bentuk lain dengan menandai [USERNAME] dan [URL]. Ini memudahkan proses pra-pemrosesan. Kedua kata kunci ini juga dihapus dalam penelitian ini karena tidak memiliki relevansi dan pengaruh yang signifikan terhadap penentuan emosi berdasarkan teks. Proses dan hasil penerapan pembersihan pada dataset yang digunakan dalam penelitian ini ditunjukkan dalam Tabel 2.

Tabel 2. Hasil Cleaning

Input	Output
Soal jln Jatibaru,polisi tdk bs GERTAK gubernur .Emangny polisi tdk ikut pmbhasan? Jgn berpolitik. Pengaturan wilayah,hak gubernur. Persoalan Tn Abang soal turun temurun.Pelik.Perlu kesabaran. [USERNAME] [USERNAME] [URL]	Soal jln Jatibaru polisi tdk bs GERTAK gubernur Emangny polisi tdk ikut pmbhasan Jgn berpolitik Pengaturan wilayah hak gubernur Persoalan Tn Abang soal turun temurun Pelik Perlu kesabaran
…	….
Ya Allah hanya Engkau yang mengetahui rasa sakit di hati ini Sembuhkanlah Ya Allah	Ya Allah hanya Engkau yang mengetahui rasa sakit di hati ini Sembuhkanlah Ya Allah

2.2.2 Case Folding

Case folding adalah tahap di mana semua huruf kapital dalam data akan diubah menjadi huruf kecil. Proses case folding ini bertujuan untuk mempermudah proses analisis dataset. Aplikasi dan hasil tahap case folding dalam dataset yang digunakan dalam penelitian ini ditunjukkan pada Tabel 3.

Tabel 3. Hasil Case folding

Input	Output
Soal jln Jatibaru polisi tdk bs GERTAK gubernur Emangny polisi tdk ikut pmbhasan Jgn berpolitik Pengaturan wilayah hak gubernur Persoalan Tn Abang soal turun temurun Pelik Perlu kesabaran	soal jln jatibaru polisi tdk bs gertak gubernur emangny polisi tdk ikut pmbhasan jgn berpolitik pengaturan wilayah hak gubernur persoalan tn abang soal turun temurun pelik perlu kesabaran

…	….
Ya Allah hanya Engkau yang mengetahui rasa sakit di hati ini Sembuhkanlah Ya Allah	ya allah hanya engkau yang mengetahui rasa sakit di hati ini sembuhkanlah ya allah

2.2.3 Tokenizing

Tokenizing adalah proses memisahkan data teks yang awalnya dalam bentuk kalimat menjadi bentuk kata-per-kata. Proses ini memiliki berbagai tujuan. Salah satunya adalah untuk memfasilitasi proses klasifikasi dan proses normalisasi. Contoh proses dan hasil aplikasi dari tahap tokenisasi dapat dilihat pada Tabel 4.

Tabel 4. Hasil Tokenizing

Input	Output
soal jln jatibarupolisi tdk bs gertak gubernur emangny polisi tdk ikut pmbhasan jgn berpolitik pengaturan wilayahhak gubernur persoalan tn abang soal turun temurunpelikperlu kesabaran username username url	soal, jln, jatibaru, polisi, tdk, bs, gertak, gubernur, emangny, polisi, tdk, ikut, pmbhasan, jgn, berpolitik, pengaturan, wilayah, hak, gubernur, persoalan, tn, abang, soal, turun, temurun, pelik, perlu, kesabaran
…	….
ya allah hanya engkau yang mengetahui rasa sakit di hati ini sembuhkanlah ya allah	ya, allah, hanya, engkau, yang, mengetahui, rasa, sakit, di, hati, ini, sembuhkanlah, ya, allah

2.2.4 Normalization

Normalization adalah proses mengubah kata-kata yang awalnya tidak standar atau tidak sempurna menjadi standar. Contoh kata nonstandar termasuk kata yang sebenarnya sudah distandarisasi tetapi disingkat dan kata yang mengandung banyak karakter berulang. Proses ini dilakukan karena karakteristik terbatas dari tweet Twitter, sehingga pengguna sering kali menggunakan teks yang pendek. Selain itu, salah satu bentuk ekspresi pengguna dalam teks adalah pengulangan banyak karakter. Penggunaan kata-kata nonstandar dan bahasa yang tidak terstruktur membuat proses ini menjadi perlu. Proses dan hasil tahap normalisasi pada dataset ini dapat dilihat pada Tabel 5.

Tabel 5. Hasil Normalization

Input	Output
soal, jln, jatibaru, polisi, tdk, bs, gertak, gubernur, emangny, polisi, tdk, ikut, pmbhasan, jgn, berpolitik, pengaturan, wilayah, hak, gubernur, persoalan, tn, abang, soal, turun, temurun, pelik, perlu, kesabaran	soal, jalan, jatibaru, polisi, tidak, bisa, gertak, gubernur, emangny, polisi, tidak, ikut, pmbhasan jangan, berpolitik, pengaturan, wilayah, hak, gubernur, persoalan, tn, abang, soal, turun, temurun, pelik, perlu, kesabaran,
….	….

ya, allah, hanya, engkau, yang, mengetahui, rasa, sakit, di, hati, ini, sembuhkanlah, ya, allah	ya, allah, hanya, engkau, yang, mengetahui, rasa, sakit, di, hati, ini, sembuhkanlah, ya, allah

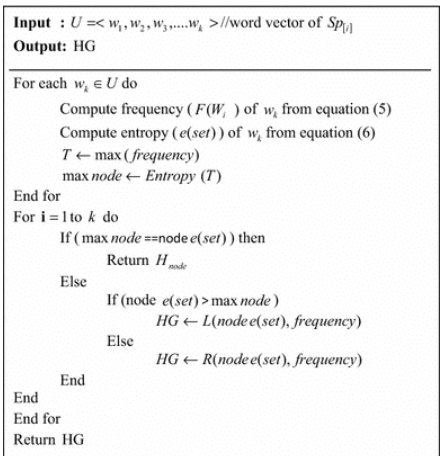
2.3. Data Split

Data Split dilakukan dengan membagi data menjadi dua bagian. Yaitu data pelatihan dan data pengujian. Dalam dataset yang bernama indonesian-twitter-emotion-dataset, data masih dalam bentuk lengkap atau belum dibagi. Oleh karena itu, tahap pembagian data pada dataset perlu dilakukan untuk mempermudah proses klasifikasi. Pembagian data dilakukan setelah data melewati tahap praproses. Menentukan proporsi data pelatihan terhadap data uji dilakukan dengan melakukan eksperimen berdasarkan penelitian terkait. Eksperimen pembagian data dilakukan dengan rasio 70:30, 80:20, dan 90:10. Setelah eksperimen, hasil tertinggi diperoleh pada rasio 70:30. Oleh karena itu, penelitian ini menggunakan rasio utama 70:30 dengan pembagian data pelatihan sebesar 70% dan data uji sebesar 30%. Pembagian dataset dilakukan menggunakan metode train_test_split dari pustaka sklearn.

2.4. Data Training

2.4.1 Hashing Vectorizer

Hashing vectorizer adalah salah satu teknik yang digunakan untuk mengubah teks menjadi bentuk vektor menggunakan algoritma hashing atau sering disebut sebagai trik hashing. Algoritma ini dapat diterapkan pada kalimat-kalimat dalam sebuah dokumen [12]. Teknik hashing vectorizer ini memiliki beberapa keunggulan dibandingkan teknik vektorisasi kata serupa. Salah satunya adalah bahwa teknik ini lebih efisien dalam penggunaan memori, yang sesuai dengan pertumbuhan ukuran data yang semakin besar [13], [14]. Langkah dalam proses mengubah teks menjadi vektor dalam teknik ini dimulai dengan mendefinisikan ukuran vektor dalam algoritma. Setelah itu, vektor yang sesuai dengan teks akan menjadi output. Teknik ini tidak terbatas pada klasifikasi teks saja, tetapi juga dapat digunakan pada tingkat dokumen [15]. Pseudocode dari algoritma hashing dapat dilihat pada Gambar 2.



Gambar 2. Pseudocode hash trick
(Source: [16])

2.4.2 Logistic Regression

Regresi Logistik atau Logistic Regression adalah salah satu algoritma analisis prediktif yang digunakan dalam masalah klasifikasi. Ini didasarkan pada konsep probabilitas dalam statistik. Regresi logistik adalah bentuk regresi linier yang lebih berkembang. Perbedaan terbesar dari regresi linier adalah bahwa titik data tidak dibentuk menjadi sebuah garis [17]. Regresi logistik itu sendiri memiliki tiga jenis termasuk regresi logistik biner, regresi logistik multinomial, dan regresi logistik ordinal [18]. Regresi logistik menggunakan fungsi biaya yang lebih kompleks yang disebut fungsi sigmoid. Hipotesis dari Regresi Logistik cenderung membatasi fungsi biaya antara 0 dan 1 dalam Regresi Logistik biner [19]. Persamaan dasar dalam algoritma regresi logistik dapat dilihat pada Persamaan 1. Persamaan fungsi sigmoid dapat dilihat pada Persamaan 2.

$$P(Y|X) = \frac{exp[\beta_0 + \sum_i \beta_i X_i]}{1 + exp[\beta_0 + \sum_i \beta_i X_i]} \quad (1)$$

Dimana:
$P(Y|X)$: Probability of Y on observation X
β_0 : Bias
β_1 : Weight vector
X_i : Observation variable

$$P = \frac{1}{1 + e^{-y}} \quad (2)$$

Dimana:
P : coefficient (*weight*)
e : constant
y : independent variable

2.4.3 Variance Threshold

Variance Threshold adalah salah satu pendekatan dasar dan sederhana untuk pemilihan fitur dalam penambangan data. Pemilihan fitur ini bekerja dengan menghapus semua fitur yang variansnya tidak memenuhi ambang batas tertentu. Penentuan nilai ambang (T) diperoleh dari hasil varians fitur. Persamaan untuk varians dapat dilihat pada Persamaan 3.

$$variance\ score(fi) = p(1 - p) \quad (3)$$

Dimana:
p : Persentase pengambilan sampel nilai fitur

Teknik ini menghapus semua fitur dengan varians nol, yaitu fitur yang memiliki nilai yang sama di semua sampel [20], [21]. Ini juga menghapus fitur dengan skor varians di bawah ambang batas. Tujuan dari penghapusan ini adalah untuk menghilangkan fitur yang memiliki variasi sangat sedikit atau yang hanya terdiri dari kebisingan.

Penyaringan ini cocok untuk data yang memiliki fitur yang diukur pada skala yang sama [22].

3. HASIL DAN PEMBAHASAN

3.1. Hasil Hashing Vectorizer

Tahap vektorisasi kata ini juga sering disebut dengan pembobotan kata. Pada penelitian ini, teknik yang digunakan adalah hashing vectorizer. Metode ini dipilih karena mampu memetakan sebuah teks ke dalam sebuah matriks kemunculan yang berisi 'nilai' tanpa mengirimkan kata yang telah diubah. Hal ini membuatnya lebih efisien dalam aplikasi dataset yang besar. Nilai hasil pada tahap ini didapatkan dengan membagi data dari proses preprocessing terlebih dahulu. Pembagian data terdiri dari data latih dan data uji dengan proporsi pembagian 70% untuk data latih dan 30% untuk data uji. Pembagian tersebut berdasarkan uji coba yang telah dilakukan oleh peneliti berdasarkan penelitian yang telah dilakukan. Data yang sebelumnya berbentuk string, diubah menjadi matriks numerik. Kemudian dilakukan tahap perhitungan dengan hashing vectorizer. Hasil dari proses pada tahap ini dapat dilihat pada tabel 6 di bawah ini.

Tabel 6. Hasil Hashing Vectorizer

Result of top hashing vectorizer		Result of bottom hashing vectorizer	
(0, 4658)	-0.18257418583505536	(3079, 14138)	0.12216944435630522
(0, 9135)	-0.3651483716701107	(3079, 14199)	-0.12216944435630522
(0, 9465)	0.18257418583505536	(3079, 16756)	0.12216944435630522
(0, 10851)	-0.18257418583505536	(3079, 17108)	0.24433888871261045
(0, 12163)	-0.18257418583505536	(3079, 19871)	0.24433888871261045
(0, 12299)	-0.18257418583505536	(3079, 26050)	-0.12216944435630522
(0, 14138)	0.3651483716701107	(3079, 28269)	-0.12216944435630522
(0, 19309)	0.18257418583505536	(3079, 28406)	-0.24433888871261045
(0, 19500)	0.18257418583505536	(3079, 31021)	-0.24433888871261045
(0, 23313)	0.18257418583505536	(3079, 36809)	0.12216944435630522
(0, 24619)	0.18257418583505536	(3079, 37966)	-0.24433888871261045
(0, 26539)	0.18257418583505536	(3079, 45226)	0.12216944435630522
(0, 28270)	0.18257418583505536	(3079, 47522)	0.12216944435630522
(0, 32923)	0.18257418583505536	(3079, 48789)	0.12216944435630522
(0, 35191)	-0.18257418583505536	(3079, 49174)	-0.12216944435630522
(0, 37196)	-0.18257418583505536	(3079, 50978)	-0.12216944435630522
(0, 43921)	-0.18257418583505536	(3079, 51298)	0.12216944435630522
(0, 45226)	0.18257418583505536	(3079, 53038)	0.12216944435630522
(0, 47194)	-0.18257418583505536	(3079, 53246)	-0.12216944435630522
(0, 58209)	0.18257418583505536	(3079, 53383)	0.12216944435630522
(0, 60487)	0.18257418583505536	(3079, 53623)	-0.12216944435630522
(0, 61138)	-0.18257418583505536	(3079, 54919)	-0.12216944435630522
(0, 62931)	0.18257418583505536	(3079, 56367)	-0.12216944435630522

3.2. Hasil Variance Threshold

Tahap seleksi fitur merupakan tahap yang dilakukan dalam analisis sentimen. Hal ini dikarenakan data hasil preprocessing dan word vectorization masih memiliki nilai fitur yang terlalu besar sehingga membuat hasil klasifikasi menjadi kurang optimal.

Dengan menggunakan variance threshold sebagai seleksi fitur, maka atribut-atribut akan disaring berdasarkan nilai variance yang telah dilakukan. Pada tahap ini digunakan library paket sklearn. Hasil penerapan seleksi fitur dengan teknik variance threshold ditunjukkan pada Tabel 7.

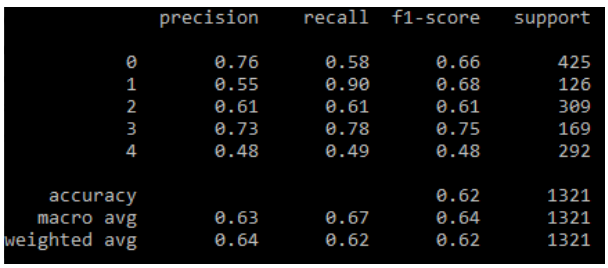
Tabel 7. Hasil Variance Threshold

Result of top feature selection		**Result of bottom feature selection**	
(0, 953)	-0,18257418583505536	(3079, 2972)	0.12216944435630522
(0, 1907)	-0,3651483716701107	(3079, 2984)	-0.12216944435630522
(0, 1988)	0,18257418583505536	(3079, 3486)	0.12216944435630522
(0, 2267)	-0,18257418583505536	(3079, 3548)	0.24433888871261045
(0, 2531)	-0,18257418583505536	(3079, 4088)	0.24433888871261045
(0, 2569)	-0,18257418583505536	(3079, 5362)	-0.12216944435630522
(0, 2972)	0,3651483716701107	(3079, 5806)	-0.12216944435630522
(0, 3985)	0.18257418583505536	(3079, 5838)	-0.24433888871261045
(0, 4020)	0.18257418583505536	(3079, 6344)	-0.24433888871261045
(0, 4771)	0.18257418583505536	(3079, 7483)	0.12216944435630522
(0, 5063)	0.18257418583505536	(3079, 7721)	-0.24433888871261045
(0, 5452)	0.18257418583505536	(3079, 9214)	0.12216944435630522
(0, 5807)	0.18257418583505536	(3079, 9655)	0.12216944435630522
(0, 6716)	0.18257418583505536	(3079, 9925)	0.12216944435630522
(0, 7140)	-0.18257418583505536	(3079, 9995)	-0.12216944435630522
(0, 7556)	-0.18257418583505536	(3079, 10362)	-0.12216944435630522
(0, 8927)	-0.18257418583505536	(3079, 10418)	0.12216944435630522
(0, 9214)	0.18257418583505536	(3079, 10770)	0.12216944435630522
(0, 9585)	-0.18257418583505536	(3079, 10822)	-0.12216944435630522
(0, 11823)	0.18257418583505536	(3079, 10849)	0.12216944435630522
(0, 12335)	0.18257418583505536	(3079, 10900)	-0.12216944435630522
(0, 12455)	-0.18257418583505536	(3079, 11168)	-0.12216944435630522
(0, 12810)	0.18257418583505536	(3079, 11455)	-0.12216944435630522

3.4 Clasiffication Report

Setelah melakukan tahap preprocessing, vektorisasi kata, dan seleksi fitur, proses selanjutnya adalah melakukan proses klasifikasi emosi berbasis teks berdasarkan algoritma klasifikasi Regresi Logistik. Untuk menghitung nilai precision, recall, f1-score, dan akurasi dari algoritma Regresi Logistik, digunakan classification report dan confusion matrix yang diimplementasikan secara terpisah. Classification report merupakan teknik untuk menentukan nilai performa dari sebuah model machine learning. Teknik ini menampilkan dalam bentuk angka yang berisi berbagai macam statistik antara lain precision, recall, f1-score dan akurasi.

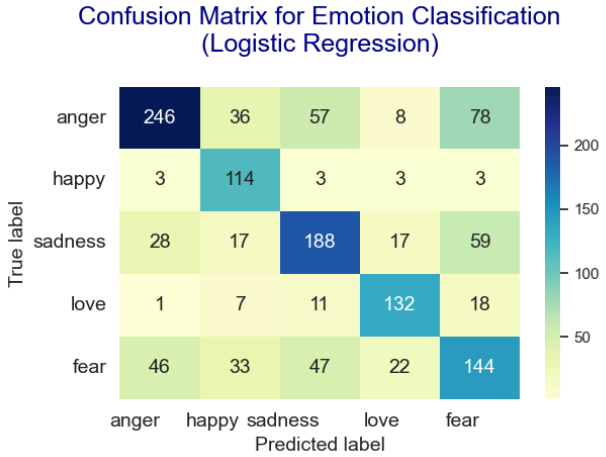
Pada classification report, penerapan hashing vectorizer dan variance threshold pada algoritma Logistic Regression menghasilkan akurasi sebesar 0.62 atau dalam satuan persen yaitu 62%. Dengan nilai precision dan recall masing-masing sebesar 0.64 atau 64% dan 0.62 atau 62%. Sementara itu, nilai F1 pada model ini adalah 0.62 atau 62%. Tampilan hasil laporan klasifikasi ditunjukkan pada Gambar 3.



	precision	recall	f1-score	support
0	0.76	0.58	0.66	425
1	0.55	0.90	0.68	126
2	0.61	0.61	0.61	309
3	0.73	0.78	0.75	169
4	0.48	0.49	0.48	292
accuracy			0.62	1321
macro avg	0.63	0.67	0.64	1321
weighted avg	0.64	0.62	0.62	1321

Gambar 3. Hasil Classification Report

Confusion matrix merupakan salah satu teknik yang dapat menentukan performa dari sebuah model dari machine learning. Performa yang ditampilkan dalam confusion



matrix adalah jumlah data dari dataset yang berhasil diprediksi dengan benar atau salah oleh model yang dibuat. Berbeda dengan classification matrix yang hanya menampilkannya dalam bentuk angka, confusion matrix ditampilkan dalam bentuk diagram plot heatmap. Hasil dari algoritma klasifikasi yang diukur menggunakan confusion matrix dapat dilihat pada Gambar 4.

Gambar 4. Hasil Confusion Matrix

Hasil dari confusion matrix menunjukkan jumlah hasil yang dapat diprediksi oleh model yang dibuat oleh peneliti. Untuk kelas marah didapatkan hasil tebakan yang sesuai dengan kelasnya dengan nilai 246 data yang dapat diprediksi dengan benar oleh model. Sedangkan total prediksi yang salah sebanyak 179 dengan rincian yang ditebak sebagai senang (happy) sebanyak 36, yang diprediksi sedih (sad) sebanyak 57, dan yang diprediksi cinta (love) 8, serta yang terakhir takut (fear) mencapai 78 data yang salah diprediksi. Kemudian untuk kelas bahagia (happy), jumlah prediksi yang tercapai dengan benar sebanyak 114 data dan masing-masing tebakan data yang salah berjumlah 3. Untuk kelas sedih (sad), model yang diusulkan peneliti berhasil menebak dengan benar sebanyak 188 data.

Sedangkan jumlah total yang diprediksi salah oleh sistem sebanyak 121 data dengan rincian 28 data diprediksi sebagai marah, 17 tebakan masing-masing diprediksi sebagai senang dan cinta, dan diprediksi sebagai takut dengan total 59. Pada kelas cinta, model berhasil memprediksi 132 data, dengan total 37 data yang salah prediksi. Pada data yang seharusnya berada pada kelas love, diprediksi sebagai marah sebanyak 1 data, diprediksi sebagai senang sebanyak 7 data, sebagai sedih sebanyak 11 data dan diprediksi sebagai takut sebanyak 18 data. Pada kelas fear, data yang diprediksi dengan benar sebanyak 144 data. Sementara itu, sebanyak 148 data salah diprediksi oleh model. Setiap kesalahan pada kelas fear diprediksi sebagai marah sebanyak 46 data, diprediksi sebagai senang sebanyak 33. diprediksi sebagai kesedihan sebanyak 47 data dan diprediksi sebagai cinta sebanyak 22 data.

3.5 Pembahasan

Penelitian ini merupakan pengembangan yang didasarkan pada beberapa referensi dimana metode, algoritma yang digunakan dan objek penelitian yang terkait atau serupa. Referensi ini dimaksudkan untuk dapat mengetahui pengembangan yang telah dilakukan, yaitu dengan cara membandingkan dengan hasil penelitian sebelumnya dengan menggunakan objek penelitian yang sama. Untuk perbandingan hasil pengukuran nilai akurasi model yang diusulkan dengan penelitian sebelumnya dapat dilihat pada Tabel 8.

Tabel 8. Perbandingan akurasi dengan penelitian sebelumnya

Researcher	***Dataset***	***Word Vectorization***	***Features***	***Classifier***	***Accuracy***
[6]	*Indonesian-Twitter-Emotion-Dataset*	*TF-IDF*	*Emotion word's list (EW)*	*Logistic Regression*	57.85%
			Bag-of-Words (BOW)	*Logistic Regression*	69.53%
			Word2Vec (WV)	*Logistic Regression*	67.32%
			FastText (FT)	*Logistic Regression*	66.46%
			EW + BOW + WV Combination	*Logistic Regression*	70.34%
			EW + BOW + FT Combination	*Logistic Regression*	73.72%
[23]	*Indonesian-Twitter-Emotion-Dataset*	*TF-IDF*	-	*Logistic Regression*	64%
				K Nearest Neighbor	49%
Proposed	***Indonesian-Twitter-Emotion-Dataset*)**	***Hashing Vectorizer***	***Variance Threshold***	***Logistic Regression***	**62%**

Pada penelitian sebelumnya yang dilakukan oleh dalam melakukan klasifikasi emosi berdasarkan teks tweet twitter bahasa Indonesia, berbagai teknik pemilihan fitur digunakan diantaranya emotion word list, bag-of-words, word2vec dan fasttext dan kombinasi fitur. Dengan menggunakan teknik tf-idf sebagai vektorisasi kata. Sedangkan tiga algoritma klasifikasi yang digunakan, salah satunya adalah regresi logistik. Kesimpulan dari penelitian ini menunjukkan bahwa algoritma regresi logistik dengan kombinasi tiga fitur dari daftar kata emosi, bag-of-words dan fasttext memiliki performa yang lebih baik dengan tingkat akurasi sebesar 73.72%. Penelitian yang dilakukan oleh [23] dalam mengklasifikasikan emosi berdasarkan teks tweet twitter menggunakan regresi logistik dan k-nearest neighbor (KNN) menunjukkan bahwa regresi logistik dapat memberikan hasil yang lebih baik.

Perbandingan dengan metode yang diusulkan pada penelitian ini untuk mengklasifikasikan emosi berdasarkan teks tweet twitter dengan vektorisasi kata menggunakan vektorisasi hashing dan dengan menggunakan seleksi fitur variance threshold menghasilkan akurasi dengan nilai 62%. Dengan melakukan perbandingan dengan metode yang telah dilakukan pada penelitian sebelumnya, menunjukkan bahwa word vectorization hashing vectorizer dan variance threshold sebagai seleksi fitur mampu memberikan nilai akurasi yang lebih baik dibandingkan dengan menggunakan seleksi fitur emotion word's list (EW) dan regresi logistik yang memiliki nilai akurasi sebesar 57.85%. Namun, masih belum dapat menghasilkan nilai akurasi yang lebih tinggi dibandingkan dengan menggunakan teknik seleksi fitur yang lain.

4. SIMPULAN

Proses kerja klasifikasi emosi berbasis teks dalam bahasa Indonesia dimulai dengan proses pra-pemrosesan, yaitu pembersihan data, case folding, tokenisasi, normalisasi, dan pembagian data. Dataset yang digunakan adalah indonesian-twitter-emotion-dataset sebelum menerapkan metode yang diusulkan dengan membaginya menjadi dua jenis data, yaitu data pelatihan dan data pengujian. Pembagian dilakukan dengan rasio 70:30. Data pelatihan diproses menggunakan algoritma regresi logistik dengan dataset yang diterapkan dalam dua tahap. Yaitu vektorisasi kata menggunakan teknik vektorisasi hashing dan pemilihan fitur menggunakan teknik ambang varians. Pada tahap vektorisasi kata, dalam penelitian ini, nilai n_features diterapkan dengan teknik vektorisasi hashing dengan nilai matriks 216. Selanjutnya, data pelatihan yang telah diproses kemudian diterapkan ambang varians yang bertujuan untuk mengurangi nilai-nilai yang tidak digunakan dalam proses klasifikasi tetapi dapat mempengaruhi nilai akurasi model yang disusun. Setelah dataset diproses menggunakan kedua teknik ini, algoritma regresi logistik kemudian disusun dan dataset yang telah diproses dimasukkan ke dalam model untuk mengukur akurasi model klasifikasi. Kemudian model yang telah disusun diuji menggunakan data uji yang kemudian dievaluasi menggunakan teknik laporan klasifikasi dan matriks kebingungan. Nilai akurasi dalam analisis menggunakan algoritma regresi logistik dengan penerapan hashing vektorisasi dan ambang varians adalah 62%. Nilai akurasi dari metode yang diusulkan dalam penelitian ini mampu menghasilkan nilai yang lebih tinggi dibandingkan dengan salah satu metode pemilihan fitur, yaitu daftar kata emosi. Namun, metode ini belum mampu menghasilkan hasil yang lebih tinggi dibandingkan dengan metode pemilihan fitur lainnya.

5. REFERENSI

[1] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and I. Awan, “Detection and classification of social media-based extremist affiliations using sentiment analysis techniques,” *Human-centric Comput. Inf. Sci.*, vol. 9, no. 1, p. 24, Dec. 2019, doi: 10.1186/s13673-019-0185-6.

[2] S. Kemp, "DIGITAL 2021: INDONESIA." [Online]. Available: https://datareportal.com/reports/digital-2021-indonesia

[3] F. Koto and G. Y. Rahmaningtyas, "Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs," in *2017 International Conference on Asian Language Processing (IALP)*, IEEE, Dec. 2017, pp. 391–394. doi: 10.1109/IALP.2017.8300625.

[4] H. A. Santoso, E. H. Rachmawanto, and U. Hidayati, "Fake Twitter Account Classification of Fake News Spreading Using Naïve Bayes," *Sci. J. Informatics*, vol. 7, no. 2, 2020, [Online]. Available: https://journal.unnes.ac.id/nju/sji/article/view/25747

[5] R. J. R. Raj, P. Das, and P. Sahu, "Emotion Classification on Twitter Data Using Word Embedding and Lexicon Based Approach," in *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, IEEE, Apr. 2020, pp. 150–154. doi: 10.1109/CSNT48778.2020.9115750.

[6] M. S. Saputri, R. Mahendra, and M. Adriani, "Emotion Classification on Indonesian Twitter Dataset," in *2018 International Conference on Asian Language Processing (IALP)*, IEEE, Nov. 2018, pp. 90–95. doi: 10.1109/IALP.2018.8629262.

[7] W. G. Parrott, *Emotions in Social Psychology: Essential Readings*. Psychology Press, 2001.

[8] T. Daouas and H. Lejmi, "Emotions recognition in an intelligent elearning environment," *Interact. Learn. Environ.*, vol. 26, no. 8, pp. 991–1009, Nov. 2018, doi: 10.1080/10494820.2018.1427114.

[9] N. Gupta, M. Gilbert, and G. Di Fabbrizio, "EMOTION DETECTION IN EMAIL CUSTOMER CARE," *Comput. Intell.*, vol. 29, no. 3, pp. 489–505, Aug. 2013, doi: 10.1111/j.1467-8640.2012.00454.x.

[10] F. Haque, M. M. H. Manik, and M. M. A. Hashem, "Opinion Mining from Bangla and Phonetic Bangla Reviews Using Vectorization Methods," in *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, IEEE, Dec. 2019, pp. 1–6. doi: 10.1109/EICT48899.2019.9068834.

[11] Y. Siti Ambarwati and S. Uyun, "Feature Selection on Magelang Duck Egg Candling Image Using Variance Threshold Method," in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, Dec. 2020, pp. 694–699. doi: 10.1109/ISRITI51436.2020.9315486.

[12] S. Gadde, A. Lakshmanarao, and S. Satyanarayana, "SMS Spam Detection using Machine Learning and Deep Learning Techniques," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, Mar. 2021, pp. 358–362. doi: 10.1109/ICACCS51430.2021.9441783.

[13] S. Kaur, P. Kumar, and P. Kumaraguru, "Automating fake news detection system using multi-level voting model," *Soft Comput.*, vol. 24, no. 12, pp. 9049–9069, Jun. 2020, doi: 10.1007/s00500-019-04436-y.

[14] N. Kumar, A. Harikrishnan, and R. Sridhar, "Hash Vectorizer Based Movie Genre Identification," 2020, pp. 798–804. doi: 10.1007/978-3-030-30577-2_71.

[15] V. Gangadharan, D. Gupta, A. L., and A. T.A., "Paraphrase Detection Using Deep Neural Network Based Word Embedding Techniques," in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, IEEE, Jun. 2020, pp. 517–521. doi: 10.1109/ICOEI48184.2020.9142877.

[16] T. Prasanth and M. Gunasekaran, "A mutual refinement technique for big data retrieval using hash tag graph," *Cluster Comput.*, vol. 22, pp. 2027–2037, 2017.

[17] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augment. Hum. Res.*, vol. 5, no. 1, p. 12, Dec. 2020, doi: 10.1007/s41133-020-00032-0.

[18] Y. Tampil, H. Komaliq, and Y. Langi, "Analisis Regresi Logistik Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Mahasiswa FMIPA Universitas Sam Ratulangi Manado," *d'CARTESIAN*, vol. 6, no. 2, p. 56, Aug. 2017, doi: 10.35799/dc.6.2.2017.17023.

[19] A. Poornima and K. S. Priya, "A Comparative Sentiment Analysis Of Sentence Embedding Using Machine Learning Techniques," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, Mar. 2020, pp. 493–496. doi: 10.1109/ICACCS48705.2020.9074312.

[20] L. Buitinck *et al.*, "API design for machine learning software: experiences from the scikit-learn project," Sep. 2013.

[21] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," Jan. 2012.

[22] D. Agustin, R. D. Atmaja, and Azizah, "Pengolahan Citra Digital untuk Mengklasifikasi Golongan Kendaraan dengan Metode Parameter Dasar Geometrik," *E-Proceeding Eng.*, no. 1, pp. 115–123, 2017.

[23] A. B. P. Negara, H. Muhardi, and F. Sajid, "Perbandingan Algoritma Klasifikasi terhadap Emosi Tweet Berbahasa Indonesia," *J. Edukasi dan Penelit. Inform.*, vol. 7, no. 2, p. 242, Aug. 2021, doi: 10.26418/jp.v7i2.48198.