Analisis Sentimen Twitter untuk Memprediksi Emosi Manusia Menggunakan Model Logistic Regression

Josha Adrista Harianto
Departemen Sistem Informasi
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
joshaharianto.205026@mhs.its.ac.id

Ichlasul Hasanat
Departemen Teknik Informatika
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ichlasulhasanat.205025@mhs.its.ac.id

Kurniaman Andreas Zega
Departemen Sistem Informasi
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
5026211056@mhs.its.ac.id

Mirza Aditya Badarudin
Departemen Sistem Informasi
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
mirzabadarudin.2050261@mhs.its.ac.id

***Abstract*— Emosi manusia adalah sebuah aspek penting dalam mempelajari sentimen sosial dan pemahaman perilaku manusia dalam berbagai medium, salah satunya merupakan media sosial seperti Twitter. Meski terlihat mudah, proses untuk memahami dan memprediksi emosi manusia merupakan sebuah proses yang sulit dan membutuhkan waktu yang lama. Penelitian ini mengusulkan penerapan berbagai metode *machine learning* untuk mengurangi usaha yang diperlukan untuk mendapatkan hasil. Data Twitter yang disediakan panitia kami gunakan sebagai sumber awal untuk analisis sentimen. Kami membandingkan performa berbagai metode *machine learning*, termasuk Logistic Regression, Gaussian Naïve-Bayes, SVC, K-Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, dan Ada Boost Classifier dalam memprediksi klasifikasi emosi manusia yang ada di dalam dataset tersebut. Penelitian ini dilakukan dengan mengukur akurasi untuk setiap metode, serta analisis perbandingan terhadap pendekatan yang sudah ada. Hasil penelitian ini memberikan pembaca pengetahuan mengenai metode ML yang paling sesuai untuk memprediksi mengklasifikasi sentimen di dalam tulisan yang ditulis manusia. Selanjutnya, informasi ini dapat diolah lebih lanjut untuk berbagai keperluan sesuai kebutuhan masing-masing.**

Keywords— Prediksi Emosi, Sentimen, Machine Learning, Twitter, Analisis Sentimen, Akurasi.

I. PENDAHULUAN

A. Latar Belakang

Di era digital yang semakin berkembang saat ini, media sosial telah menjadi salah satu *platform* utama di mana orang-orang berbagi pemikiran, perasaan, dan pendapat mereka. Twitter, sebagai salah satu platform media sosial yang paling populer, telah menjadi wadah bagi pengguna untuk mengekspresikan berbagai emosi dan perasaan mereka secara publik. Menurut data yang dirilis oleh situs Data Reporters, pada April 2023, jumlah pengguna Twitter di Indonesia mencapai 14,75 juta orang, menempatkannya sebagai negara dengan peringkat keenam terbesar di dunia dalam hal jumlah pengguna Twitter (Reporter, Data, 2023). Data Reporter juga mengungkapkan bahwa Indonesia merupakan negara ketiga terbanyak di dunia dalam hal jumlah tweet yang diposting, dengan kontribusi sekitar 11,39% dari total tweet yang telah direkam sejak November 2010 yang dibuat oleh 383 juta profil pengguna Twitter yang dibuat sebelum tanggal 1 Januari 2012. Hal ini menunjukkan bahwa pasar Twitter di Indonesia memiliki potensi besar, yang mendorong berbagai produsen, baik skala kecil maupun besar, untuk bersaing guna memanfaatkan potensi ekonomi yang besar ini.

Analisis sentimen, yang merupakan bagian dari opinion mining, telah menjadi target utama dalam mengidentifikasi dan memahami perilaku manusia, terutama di media sosial. Proses menganalisis ini, apabila dilakukan secara manual akan membutuhkan sangat lama karena banyaknya hal yang perlu diperhatikan. Oleh akrena itu, kami menggunakan *machine learning* untuk dapat memproses data dalam jumlah besar secara cepat, efisien, dan akurat. Peneltiian ini menggunakan metode-metode seperti Logistic Regression, Gaussian Naïve-Bayes, SVC, K-Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, dan Ada Boost Classifier untuk melihat model mana yang dapat menghasilkan akurasi yang paling tinggi.

Harapan kami, penelitian ini dapat membuka gerbang penelitian dan riset yang lebih mendalam di masa depan. Penelitian mengenai perasaan manusia kami rasa akna semakin marak untuk memberikan wawasan yang lebih mendalam mengenai cara manusia berinteraksi antara satu sama lain.

Rumusan Masalah

(1) Bagaimana cara memproses dataset yang disediakan agar siap diguakan untuk analisis?
(2) Apa metode terbaik umtuk memprediksi sentimen yang ditimbulkan user?
(3) Seberapa akurat hasil dari metode *machine learning* terbaik?

Tujuan Penelitian

(1) Mengetahui cara memproses dataset yang disediakan agar siap diguakan untuk analisis
(2) Mengetahui metode terbaik untuk melakukan prediksi emosi manusia
(3) Seberapa akurat hasil dari metode *machine learning* terbaik

Manfaat Penelitian

(1) Meningkatkan pemahaman penulis dan pembaca mengenai emosi manusia dalam konteks media sosial.
(2) Mengurangi waktu yang dibutuhkan untuk mengidentifikasi emosi manusia, terutama yang berasal dari Twitter.
(3) Menjadi dasar untuk selanjutnya mengaplikasikan pengetahuan pada penelitian ini di dunia nyata.

II. PEMBAHASAN

A. *Landasan Teori*

Analisis Sentimen

Analisis Sentimen merupakan suatu proses yang melibatkan pencarian makna dari pendapat, pandangan, atau emosi yang terdapat dalam teks, ucapan, postingan (yang ditemukan di internet), atau dalam basis data, menggunakan pendekatan Natural Language Processing (NLP) [2]. Dalam analisis sentimen, terdapat beberapa tugas yang mencakup ekstraksi sentimen, klasifikasi sentimen, dan deteksi spam [3]. Analisis sentimen memiliki peran penting di dunia bisnis karena membantu perusahaan mengetahui serta memahami respon dari pelanggan terhadap produk yang ditawarkan [4].

Pendekatan dasar dalam menganalisis sentimen dalam teks melibatkan pembuatan kamus kata, yang terbagi menjadi kata-kata positif dan kata-kata negatif, berdasarkan sifat intrinsik dari kata tersebut. Tingkat akurasi dari metode ini sangat bergantung pada sejauh mana kamus kata tersebut lengkap [5].

Aplikasi Twitter

Aplikasi twitter merupakan salah satu media sosial paling popular saat ini. Berita terbaru, hal-hal penting sampai hal yang tidak penting, menumpahkan perasaa atau opini semuanya terjadi di aplikasi ini. Twitter menyediakan platform bagi penggunanya untuk mengirim, membaca, dan membalas pesan singkat yang biasa disebut '*tweets*'. Dengan adanya kebebasan dan adanya rasa aman pengguna lebih bebas mengespresikan apa yang sedang dirasakan dan apa yang menjadi hal terkini.

Text Mining

Text Mining adalah suatu algoritma yang digunakan untuk mengekstraksi informasi dari teks yang terstruktur. Teknik ini juga memungkinkan pengguna untuk memahami konten dari sebuah teks tanpa perlu membaca seluruhnya. Pengolahan data dalam Text Mining melibatkan beberapa tahapan yang dikenal sebagai Text Preprocessing. Text Preprocessing terdiri dari beberapa langkah, termasuk Transformasi Kasus, Penghapusan Stopword, Tokenisasi, dan Stemming [6].

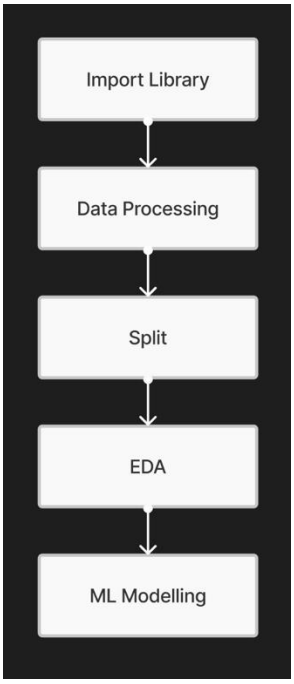
Machine Learning

Machine learning merupakan suatu cabang ilmu yang berkaitan erat dengan ilmu kecerdasan buatan. Ilmu ini merupakan suatu studi yang mempelajari mengenai algoritma serta model statistik dalam melakukan suatu *task* dalam sistem komputer, dimana sistem mampu mempelajari *task* tanpa perlu diberikan suatu perintah khusus. *Machine learning* bekerja berdasarkan pola yang diberikan dan juga bagimana kesimpulan yang bisa ditarik. Dalam mendapatkan pola dan kesimpulan, *machine learning* menggunakan data sampel atau yang biasa disebut *'training data'* sebagai suatu model matematika [7].

Logistic Regression

Meskipun disebut "regresi," *logistic regression* sebenarnya adalah model klasifikasi daripada model regresi. *Logistic regression* adalah metode yang simpel dan lebih efisien untuk menangani masalah klasifikasi biner dan linier. Logistic regression merupakan model klasifikasi yang mudah diimplementasikan dan dapat mencapai kinerja yang sangat baik ketika kelas-kelas dapat dipisahkan secara linear. Di industri, algoritma ini sering digunakan untuk tugas klasifikasi. Seperti Adaline dan perceptron, model *logistic regression* adalah metode statistik untuk klasifikasi biner yang dapat diperluas ke klasifikasi multikelas. Scikit-learn memiliki versi implementasi regresi logistik yang sangat dioptimalkan dan mendukung tugas klasifikasi multikelas [8].

B. *Metode Penelitian*



Gambar 1 Gambaran Umum Sistem

Preprocessing

Tahapan ini terdiri dari beberapa tahapan yakni cleaning, case folding, parsing, dan filtering.

1. Cleaning: Pembersihan teks mencakup penghapusan karakter khusus, tanda baca, angka, dan kata atau frasa tertentu seperti [USERNAME] dan [URL]. Tahap ini dilakukan dengan menggunakan ekspresi reguler (regex) dan pemisahan kata dengan split.
2. Case Folding: Semua huruf dalam teks diubah menjadi huruf kecil dengan menggunakan text.lower().

3. Parsing: Beberapa langkah parsing dilakukan, seperti menghapus karakter HTML dengan menggantinya dengan spasi dan menghapus karakter slain huruf dari A-Z. Ini membantu untuk membersihkan teks dari elemen HTML dan karakter non-alfabet.
4. Filtering: Kata-kata yang tidak relevan, seperti kata penghubung dan kata ganti pribadi dan kata-kata lain yang ada dalam daftar stopwords dihapus. Ini juga mencakup menghapus kata-kata atau frasa tertentu yang terdaftar dalam words_to_remove.

ML Modelling

Pada Proses ini dilakukan pemilihan algoritma model ML. Setelah dilakukan pengujian, didapatkan bahwa logistic regression mendapatkan akurasi yang cukup baik dari model lainnya dengan menggunakan optimasi *hyperparameter* CountVectorizer. Setelah memilih model yang cocok untuk penelitian ini, dilakukan penyesuaian kembali dengan menggunakan *vectorizer* lain yakni GridSearchCV. Dengan GridSearchCV, didapatkan bahwa akurasi dari model mengalami peningkatan yakni dari 0,667 ke 0,711. Pada proses ini juga dilakukan pembuatan pipeline untuk menggabungkan pemrosesan model dengan TF-IDF. TF-IDF adalah suatu metode untuk memproses teks menjadi representasi numerik yang dapat digunakan oleh model machine learning.

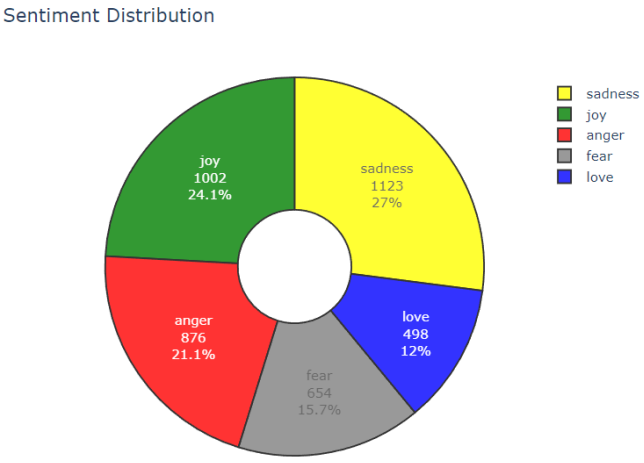
Validasi Data

Pada machine learning, ditetapkan validasi dengan menggunakan 3-fold cross validation. Dengan cross validation, data akan dibagi menjadi 3 bagian dan divalidasi 3 kali secara berulang.

III. HASIL DAN PEMBAHASAN

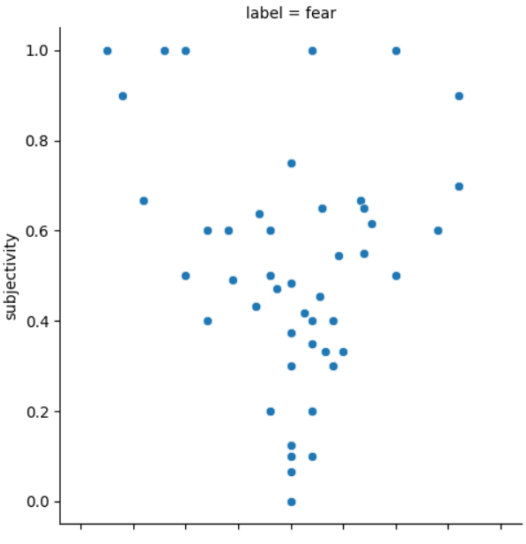
Panitia memberikan kami total 5,154 baris data berupa tweet yang ditulis oleh akun anonim di Twitter. Dari seluruh baris data tersebut, 4,153 baris data sudah diberikan label mengenai nada yang dikandung dalam setiap tweet. Sementara sisanya tidak diberikan label sama sekali. Kami melengkapi dataset ini dengan mengisi label-label yang masih kosong dalam dataset yang diberikan. Proses dimulai dengan pembersihan data dimana kami melakukan persiapan agar hanya data bersih yang akan dikenakan ML. Langkah-langkah seperti memastikan nilai dalam bentuk string, menghapus kata-kata tidak relevan, hingga menyerahamkan huruf kecil kami lakukan demi optimalisasi hasil akhir. Selain itu, masih banyak lagi pembersihan yang kami lakukan untuk membersihkan data yang akan dianalisis. Kebersihan data akan sangat mempengaruhi akurasi dari metode yang dipilih. Hal ini disebabkan data bersih akan menunjukkan informasi yang terkandung dalam dataset tanpa adanya ganggunan apapun. Sehingga memungkinkan untuk kebenaran muncul dari dalam data yang dimiliki. Kami juga menghapus baris kosong yang masih terdapat di dalam dataset yang diberikan. Pada akhirnya kami memproses 4,151 baris data dalam penelitian ini.

Dataset yang kami gunakan terdiri atas dua kolom yang berisi tentang twit yang akan dianalisis dan klasifikasi dari nada yang dibuat berdasarkan twit yang berkaitan. Untuk melihat persebaran datanya, kami telah membuat sebuah *pie chart* untuk mempermudah memahami dataset yang akan dianalisis:

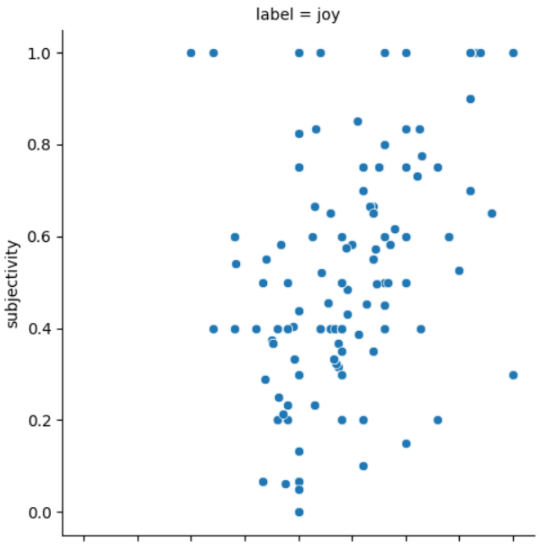


Gambar 2 Persebaran Klasifikasi Emosi Label

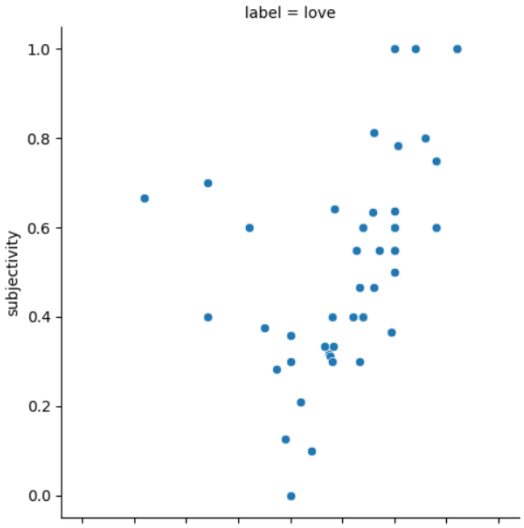
Setelah mengetahui persebaran dari setiap baris dataset, kami selanjutnya membedah masing-masing dari nada yang ditimbulkan untuk mengetahui kebenaran dari masing-masing nada yang sudah diklasifikasi. Kami menggunakan polarity dan subjectivity untuk membuktikan validitas dari twit dengan klasifikasi yang diberikan. Dalam grafik terdapat dua sumbu berbeda, yaitu sumbu x yang mewakili polarity dan sumbu y mewaliki subjectivity. Semakin mendekati 1 angka dari kedua sumbu, maka semakin besar keterkaitan antara kata-kata yang terkandung dalam tweet dengan klasifikasi yang sudah diberikan. Berikut merupakan pemaparannya:



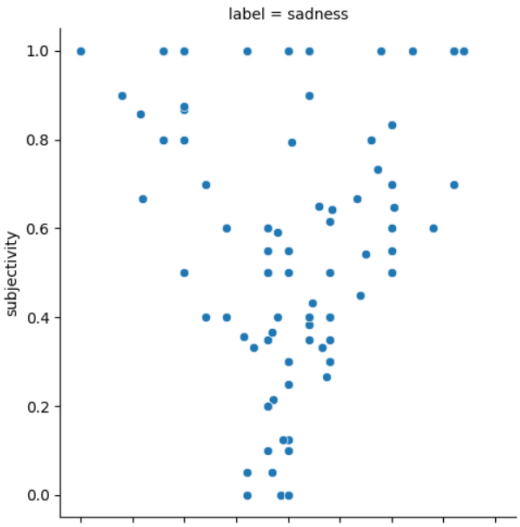
Gambar 3 Klasifikasi "fear" sebelum training



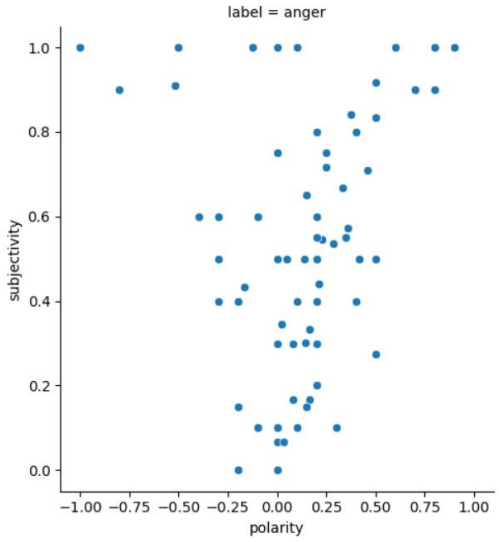
Gambar 4 Klasifikasi "joy" sebelum training



Gambar 5 Klasifikasi "love" sebelum training



Gambar 6 Klasifikasi "sadness" sebelum training



Gambar 7 Klasifikasi "anger" sebelum klasifikasi

Seluruh *labeled data* tersebut kemudian kami bagi menjadi train dan test set sebagai syarat untuk melakukan analisis. Dalam konteks analisis, pemilihan metode ML vital perannya karena setiap dataset yang digunakan memiliki karakteristik tersendiri. Kemudian, setiap metode ML juga memiliki cara kerjanya tersendiri. Merupakan tantangan tersediri bagi penulis untuk menemukan metode ML dengan akurasi yang paling tinggi untuk menjaga integritas dari dataset yang sudah diberikan. Kami menggunakan *Count Vectorizer* sebagai library untuk mengubah text data menjadi data numerical. Setelah itu, akan kami train menggunakan tujuh model ML yang berbeda dan mencari scenario terbaik yang kemudian akan kami kembangkan lebih lanjut. Berikut merupakan hasilnya:

Tabel 1: Hasil Metode Logistic Regression

Classifier	**Logistic Regression**		
Accuracy	**0.6679**		
	Precision	Recall	F1-Score
Anger	0.65	0.64	0.65
Fear	0.79	0.66	0.72
Joy	0.68	0.71	0.79
Love	0.79	0.67	0.73
Sadness	0.56	0.65	0.60

Tabel 2: Hasil Metode Gaussian Naïve-Bayes

Classifier	**Gaussian Naïve-Beyes**		
Accuracy	**0.4813**		
	Precision	Recall	F1-Score
Anger	0.52	0.65	0.58
Fear	0.49	0.41	0.45
Joy	0.52	0.54	0.53
Love	0.36	0.37	0.36
Sadness	0.45	0.39	0.42

Tabel 3: Hasil Metode Gaussian SVC

Classifier	**SVC**		
Accuracy	**0.6534**		
	Precision	Recall	F1-Score
Anger	0.59	0.64	0.61
Fear	0.76	0.69	0.72
Joy	0.65	0.66	0.65
Love	0.79	0.75	0.77
Sadness	0.59	0.60	0.59

Tabel 4: Hasil Metode K-Neighbors

Classifier	**K-Neighbors**		
Accuracy	**0.4152**		
	Precision	Recall	F1-Score
Anger	0.40	0.49	0.44
Fear	0.55	0.28	0.37
Joy	0.62	0.20	0.30
Love	0.52	0.59	0.55
Sadness	0.33	0.58	0.42

Tabel 5: Hasil Metode Decision Tree

Classifier	**Decision Tree**		
Accuracy	**0.5415**		
	Precision	Recall	F1-Score
Anger	0.51	0.50	0.50
Fear	0.75	0.61	0.67
Joy	0.50	0.51	0.51
Love	0.64	0.67	0.66
Sadness	0.45	0.49	0.47

Tabel 6: Hasil Metode Random Forest

Classifier	**Random Forest**		
Accuracy	**0.6498**		
	Precision	Recall	F1-Score
Anger	0.64	0.65	0.65
Fear	0.91	0.67	0.77
Joy	0.62	0.59	0.60
Love	0.80	0.76	0.78
Sadness	0.52	0.64	0.57

Tabel 7: Hasil Metode Ada Boost

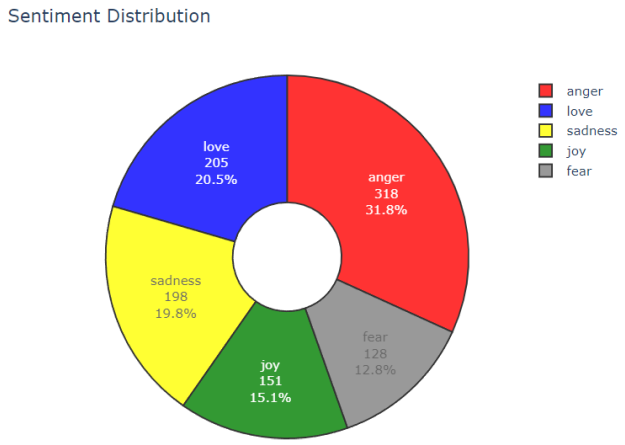
Classifier	**Ada Boost**		
Accuracy	**0.4573**		
	Precision	Recall	F1-Score
Anger	0.61	0.13	0.21
Fear	0.87	0.54	0.67
Joy	0.51	0.20	0.29
Love	0.71	0.56	0.62
Sadness	0.34	0.87	0.48

Berdasarkan hasil klasifikasi menggunakan beberapa model di atas. Didapatkan model dengan akurasi tertinggi yaitu menggunakan *Logistic Regression*, dengan akurasi sebesar 0. 6679. Namun akurasi ini masih terbilang cukup kecil. Sehingga kami memutuskan untuk mengubah library konversi text data dari *Count Vectorizer* menjadi *Grid Search Count Vectorizer*. Hasil yang didapat sudah dijelaskan di BAB II pada sub judul "ML Modelling". Berikut merupakan beberapa baris contoh hasil model setelah training:

Tabel 8: Hasil Metode Klasifikasi

No.	Tweet	Label
1	taKan raguKan besarNya kasih sayang Tuhan telah diberiKanNya kepada sayawalau momen sulit harus hadapi dalam hidup Karna saat cahaya dalam hati kita mulai reduphanya kasih Sayang dari mereka mencintai kita mampu menjadi lentera tuk menerangiNya kembali	love
2	Cc in ke cebong2 dungu Maksd hati pengen cari kesalahan lwat pohon plastik malah kebongkar semua skandalnya Sekali lagi kasian Ahok 2019GantiPresiden	sadness
3	Melody masih membatasi diri ala member ya dibales mentionnya pun dikenal aja hahaha Ya kalau seperti sih kisah Jeketi nda jauh beda dengan sebelumsebelumnya	anger
4	Rasa amarah membuatku merasa seperti akan meluapkan semuanya	Sadness
5	Rasa amarah membuatku merasa seperti akan meluapkan semuanya	anger

Terakhir, berikut merupakan *pie chart* yang menunjukkan sebaran klasifikasi yang dimiliki oleh data setelah training:



Gambar 8 Klasifikasi setelah training

IV. KESIMPULAN

1. Pemrosesan dataset agar menjadi data bersih yang siap diproses sangat dipengaruhi oleh datasetnya itu sendiri. Tidak ada dua kasus dataset yang identik. Masing-masing pasti memiliki karakteristiknya sendiri. Oleh karena itu, membutuhkan penanganan tersediri. Contohnya pada dataset ini, kami menitikberatkan pembersihan data pada karakter khusus seperti "#$@%" dan kata-kata yang tidak relevan dengan target seperti "atau", "saya", dan "yang". Jumlah data bersih yang dihasilkan melalui proses preprocessing 4,151 baris data bersih
2. Kami membandingkan tujuh metode *machine learning* berbeda untuk menentukan metode terbaik untuk dataset yang kami miliki. Setelah membandingkan performa dari seluruh metode, kami menemukan bahwa *Logistic Regression* dapat memberikan akurasi paling baik.
3. Berdasarkan hasil analisis ke-7 metode berbeda, kami menemukan bahwa metode Logistic Regression memiliki akurasi prediksi label yang paling tinggi jika dibandingkan dengan metode-metode lainnya. Metode ini mampu menghasilkan nilai akurasi sebesar 0.711 dari total nilai 1.

REFERENCES

[1] C. M. Annur, "Jumlah Pengguna Twitter di Indonesia Capai 14,75 Juta per April 2023, Peringkat Keenam Dunia," 31 May 2023. [Online]. Available: https://databoks.katadata.co.id/datapublish/2023/05/31/jumlah-Available:https://databoks.katadata.co.id/datapublish/2023/05/31/jumlah-pengguna-twitter-di-indonesia-capai-1475-juta-per-april-2023-peringkat-keenam-dunia#:~:text=Media-,Jumlah%20Pengguna%20Twitter%20di%20Indonesia%20Capai%2014%2C75%20Juta,April%202023%2C%20Pering. [Accessed 7 September 2023].J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] V. A. Flores, L. Jasa and Linawati, "Analisis Sentimen untuk Mengetahui Kelemahan dan Kelebihan Pesaing Bisnis Rumah Makan Berdasarkan Komentar Positif dan Negatif di Instagram," *Majalah Ilmiah Teknologi Elektro*, vol. 19, no. 1, 2020.

[3] V. A. Kharde and S. Sonwane, "Sentiment Analysis of Twitter Data: A Survey of Techniques," *International Journal of Computer Applications*, vol. 139, no. 11, 2016. R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[4] B. Nandi, M. Ghanti and S. Paul, "Text Based Sentiment Analysis," *Proceedings of the International Conference on Inventive Computing and Informatics*, 2017.

[5] X. Fan, "Text Sentiment Analysis: A Review," in *IEEE 4th International Conference on Computer and Communications*, Beijing, 2018.

[6] A. T. J. Harjanta, "Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining," *Jurnal Informatika Upgris*, vol. 1, no. 11, 2015.

[7] A. Subasi, Practical Machine Learning for Data Analysis Using Python, Academic Press, 2020.