



Analisis sentimen masyarakat terhadap kebijakan vaksinasi covid-19 pada media sosial twitter menggunakan metode logistic regression

Ash Shiddicky^{*1}, Surya Agustian²Email: ¹11750115325@students.uin-suska.ac.id, ²surya.agustian@uin-suska.ac.id¹ Teknik Informatika, Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau² Teknik Informatika, Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau

Diterima: 6 Juli 2022 | Direvisi: 1 Agustus 2022 | Disetujui: 13 Agustus 2022

©2020 Program Studi Teknik Informatika Fakultas Ilmu Komputer,
Universitas Muhammadiyah Riau, Indonesia

Abstrak

Kebijakan melakukan vaksinasi ini hadir sebagai upaya pemerintah dalam merespon wabah penyakit menular Covid-19, tentu dalam suatu kebijakan tidak terlepas dari dukungan maupun kontra. Berbagai komentar ini muncul dari masyarakat baik berupa opini atau fakta pengalaman yang mereka alami, ratusan bahkan ribuan komentar ini dapat menjadi data yang berharga untuk dijadikan bahan analisis dan mengetahui reaksi masyarakat ketika kebijakan vaksinasi Covid-19 ini diterapkan ke masyarakat, data komentar ini dapat diproses untuk mendapatkan informasi, salah satu analisis pendapat atau review dapat dilakukan adalah menggunakan analisis sentimen. Berdasarkan uraian diatas, dapat memanfaatkan pembelajaran mesin menggunakan algoritma Logistic Regression untuk membantu dalam mengolah data informasi tersebut, dalam hal ini klasifikasi. Hasil pengujian menunjukkan tingkat akurasi mencapai 82% terhadap data dev akan tetapi tidak diikuti oleh nilai F1-Score yang cukup baik untuk membangun sebuah model, hal ini disebabkan oleh data train memiliki distribusi klasifikasi yang tak seimbang. Maka dilakukan proses Slicing data serta Tuning Hyperparameters yang bertujuan untuk mendapatkan model terbaik, setelah menerapkan model terbaik yang didapat dari data dev maka dilakukan pengujian terhadap data test, metode Logistic Regression menghasilkan nilai akurasi 67% dan F1-score 60% terhadap data test hasil ini membuktikan bahwa model yang dibangun cukup handal dalam melakukan klasifikasi, terbukti nilai f1-score dan akurasi cukup berimbang dan lebih baik dari Naïve Bayes, SVM dan LSTM.

Kata kunci: analisis sentimen; machine learning; klasifikasi; logistic regression; f1-score; tuning hyperparameters.

Analysis of public sentiment against the covid-19 vaccination policy on twitter social media using the logistic regression method

Abstract

This vaccination policy is present as the government's effort in responding to the outbreak of the Covid-19 infectious disease, of course in a policy that cannot be separated from support and contra. These comments come from the community in the form of opinions or facts from their experiences, hundreds or even thousands of these comments can be valuable data to be used as material for analysis and to find out the public's reaction when the Covid-19 vaccination policy is applied to the community, this comment data can be processed to get information, one of the opinions analyzes or reviews that can be done is using sentiment analysis. Based on the description above, machine learning can use the Logistic Regression algorithm to assist in processing the information data, in this case classification. The test results show an accuracy rate of 82% for the dev data but it is not followed by the F1-Score value which is good enough to build a model, this is because the train data has an unbalanced classification distribution. Then the process of Slicing data and Tuning Hyperparameters is carried out which aims to get the best model, after applying the best model obtained from the dev data, a test is carried out on the test data, the Logistic Regression method produces an accuracy value of 67% and an F1-score of 60% against the test data. Proves that the built model is reliable enough to classify, it is proven that the f1-score value and accuracy are pretty balanced and better than Naïve Bayes, SVM and LSTM.

Keywords: sentiment analysis; machine learning; classification; logistic regression f1-score; tuning hyperparameters.

1. PENDAHULUAN

Vaksin adalah zat biologis yang berfungsi sebagai jenis respons imun terhadap antigen tertentu yang ditemukan dalam virus penyebab penyakit menular [1], Pelaksanaan vaksin dalam rangka penanganan pandemi Corona Virus Disease 2019 tertuang dalam Peraturan Menteri Kesehatan Nomor 10 Tahun 2021, dimana peraturan ini muncul bertujuan untuk mencapai Herd Immunity (Kekebalan kelompok) di masyarakat [2]. Penelitian yang diadakan oleh Kementerian Kesehatan Republik Indonesia bersama Indonesia Technical Advisory Group on Immunization (ITAGI) serta dukungan dari UNICEF dan organisasi kesehatan dunia WHO, menyimpulkan bahwa dua pertiga responden siap menerima vaksin, hal ini didapatkan dari tiga perempat responden yang mengaku telah mendengar vaksin covid-19, penelitian ini dilakukan pada rentang 13 sampai 19 september 2020 dimana sekitar 54% atau separuh dari populasi responden lebih menyukai platform media sosial seperti WhatsApp, Instagram, Facebook, dan Twitter untuk mempelajari lebih lanjut tentang vaksinasi Covid-19, diikuti oleh media cetak dan massa seperti televisi dan surat kabar, dan 13% lebih menyukai metode telekomunikasi terdahulu seperti SMS atau telepon [3].

Kebijakan ini tentu saja mengundang respon masyarakat sebagaimana yang dijumpai di platform media sosial seperti Twitter, hal ini didasari atas penolakan dari masyarakat akan rendahnya tingkat efikasi atau penilaian terhadap vaksin yang hanya menghasilkan nilai efikasi sebesar 65% serta diikuti oleh pernyataan anggota Dewan Perwakilan Rakyat (DPR) yang mengeluarkan pernyataan lebih memilih untuk membayar denda dibandingkan melakukan vaksin, Kebijakan ini sampai sekarang masih menuai berbagai komentar, baik berupa komentar dukungan yang bersifat positif, negatif atau komentar yang tidak berpihak kepada salah satunya yakni netral. Dalam beberapa hal, produk kebijakan politik merupakan sesuatu yang tidak dapat diganggu gugat dan tidak dapat dihindari maka [4] dalam kebijakan politik tidak dapat dihindari [5]. Kesan masyarakat terhadap upaya imunisasi Covid-19 yang miring disebabkan oleh kurangnya pengetahuan masyarakat, dan hal ini muncul sebagai akibat dari komunikasi yang buruk dari pihak berwenang, seperti petugas kesehatan, yang gagal membujuk masyarakat tentang kegunaan vaksin [6]. Berbagai komentar ini muncul dari masyarakat baik berupa opini atau fakta pengalaman yang mereka alami, ratusan bahkan ribuan komentar ini dapat menjadi data yang berharga untuk dijadikan bahan analisis dan mengetahui reaksi masyarakat ketika kebijakan vaksinasi Covid-19 ini diterapkan ke masyarakat, data komentar ini dapat diproses untuk mendapatkan informasi.

Salah satu analisis pendapat atau review dapat dilakukan adalah menggunakan analisis sentimen, analisis sentimen atau opinion mining adalah suatu studi komputasi tentang pandangan, penilaian, sikap, dan emosi orang tentang suatu hal, individu, situasi, peristiwa, topik, dan atributnya [7]. Hasil studi komputasi tersebut dapat dijadikan sebagai acuan pemerintah untuk mengambil kebijakan kedepannya baik sebagai bahan evaluasi terhadap kebijakan tersebut atau sebagai bahan pembelajaran bagi mesin. Ada banyak studi komputasi untuk mengolah sentiment diantaranya Support Vector Machine, LSTM, Decision Tree, RNN, K-Nearest Neighbors, Random Forest, Naive Bayes Classifier, juga termasuk didalamnya metode pengelompokan data atau klasifikasi data probabilitas statistik yakni Logistic Regression [8].

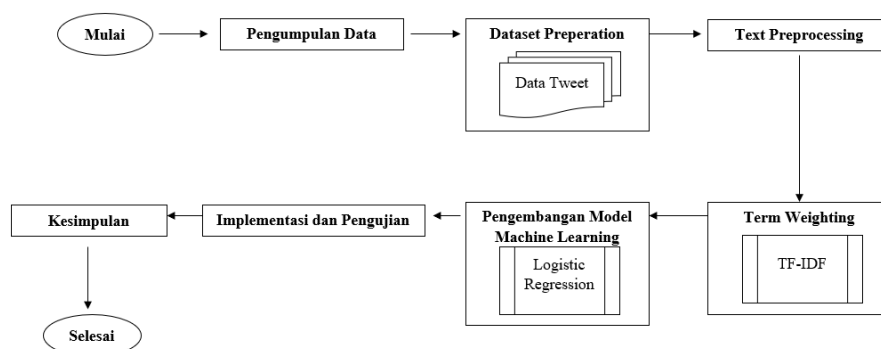
Pada penelitian sentimen analisis yang berjudul Sentiment Analysis of Azerbaijani tweets Using Logistic Regression, Naive Bayes and SVM terbukti metode Logistic Regression mampu melakukan klasifikasi tweet dengan tingkat akurasi 93 persen [9], selanjutnya dalam penelitian yang berjudul Sentiment Analysis Using Twitter Data Regarding BPJS Cost Increase and Its Effect on Health Sector Stock Prices menghasilkan kesimpulan bahwa respon sentimen masyarakat terhadap kenaikan BPJS didominasi oleh sentimen negatif yang memperlumahkan pelayanan yang kurang baik dan masalah terkait biaya, sedangkan respon positif berisi program BPJS memberi pelayanan gratis, berikutnya kebijakan ini juga berpengaruh secara signifikan pada 2 saham BUMN yakni Kimia Farma dan Indofarma ini diperoleh berdasarkan hasil pemodelan Logistic Regression yang tingkat akurasi pemodelannya mencapai 69,6 persen dan 73,9 persen [10]. Pada penelitian Dewi, dkk menggunakan metode Logistic Regression untuk klasifikasi pemilihan program studi MIPA, menghasilkan kesimpulan bahwa metode ini terbukti klasifikasi sesuai terhadap pemilihan program studi [11]. Selanjutnya dalam penelitian tentang Analisis sentimen penerapan PSBB di DKI Jakarta dan dampaknya terhadap pergerakan IHSG, metode regresi logistik menghasilkan akurasi paling tinggi yakni 75% selanjutnya diikuti oleh KNN dan Naive Bayes 70% serta Random Forest dengan hasil akurasi sebesar 60% .

Sebagaimana permasalahan yang telah diuraikan diatas maka akan dilakukan penelitian klasifikasi sentimen vaksinasi Covid-19 di Indonesia dengan menggunakan metode Logistic Regression. Tujuan dilakukan penelitian ini untuk mengklasifikasikan kalimat atau sentimen kedalam kelas positif, negatif dan kelas netral menggunakan metode Logistic Regression dan mengetahui tingkat akurasi dari metode Logistic Regression dalam mengklasifikasikan sentimen terhadap kebijakan vaksinasi Covid-19 pada media sosial Twitter.

2. METODE PENELITIAN

2.1 Tahapan Penelitian

Metodologi penelitian adalah gagasan kerangka penelitian yang disusun secara metodis sehingga saat melakukan penelitian, membahas masalah yang muncul dalam perumusan pertanyaan penelitian dan mencapai hasil yang diinginkan. Berikut ini adalah rincian langkah-langkah yang akan diselesaikan dalam rangka penyelesaian penelitian ini, seperti terlihat pada gambar di bawah ini.



Gambar 1. Metodologi Alur Penelitian

2.2 Pengumpulan Data

Pada tahap ini merupakan tahapan pengumpulan data, dimana data yang telah berhasil dihimpun dari komentar atau tweet ulasan yang berasal dari media social Twitter. Data dikumpulkan melalui proses crawling data menggunakan API yang disediakan dari Twitter yakni menggunakan library tweepy dengan menggunakan bahasa pemrograman python, kemudian dilakukan proses labelling yang bertujuan untuk memberi label atau nama atribut kedalam kumpulan dataset [12], pelabelan sentimen yang dilakukan secara crowdsourcing. Crowdsourcing sendiri merupakan metode dengan cara menghimpun tenaga orang dari luar untuk dimintai menganalisa atau menyelesaikan suatu pekerjaan yang bertujuan melakukan suatu pekerjaan demi suatu tujuan bersama. Data yang berhasil dikumpulkan sebanyak 9178 data yang berhubungan dengan kebijakan vaksin di media sosial Twitter.

2.3 Dataset Preparation

Data yang diolah merupakan data tweet ulasan di Twitter, data telah didapatkan sebanyak 9178 tweet yang di annotation menggunakan crowdsourcing membentuk 3 kelas, yakni kelas positif, negatif dan netral. Selanjutnya data tersebut dalam prosesnya dipisah menjadi data training, development serta testing. Data training merupakan jenis data tweet yang kemudian dijadikan sebagai pembelajaran pada suatu model yang ingin dibangun, berikutnya untuk memvalidasi model telah dibangun menggunakan data development, sedangkan data testing adalah data yang bertujuan untuk pengujian model sebagai benchmark atau simulasi penggunaan model pada data sebenarnya yang akan dihadapi, dataset penelitian ini pada Tabel 1. Sebagai berikut.

Tabel 1. Dataset

Dataset	Positif	Netral	Negatif	Jumlah
Train	463	6664	873	8000
Dev	45	648	85	778
Test	108	234	58	400

2.4 Teks Preprocessing

Berikutnya adalah langkah memproses data teks sehingga bisa dijadikan bahan untuk analisa serta pada langkah ini dapat menjadikan data lebih terstruktur serta mengurangi jumlah noise pada data yang didapat dari hasil Crawling Twitter. Selanjutnya akan dilakukan seleksi pemilihan fitur kombinasi terbaik yang dapat menghasilkan akurasi paling tinggi dalam metode ini. Dalam langkah ini ada beberapa proses yang umum dilakukan, yaitu:

1. Cleaning, merupakan proses membuang informasi yang dianggap kurang diperlukan atau berhubungan seperti mention, emoticon dan Uniform Resource Locator (URL)
2. Case Folding, langkah untuk mengkonversi sebuah teks untuk diubah ke bentuk yang seragam, dalam hal ini adalah lowercase yakni merubah teks menjadi huruf kecil.
3. Tokenizing, tahapan memecah suatu kalimat membentuk potong-potongan berupa kata maupun token.
4. Stopword Removal, merupakan tahapan untuk menghilangkan atau membersihkan kata-kata yang sekiranya tidak diperlukan seperti kata ganti, depan dan sambung serta kata-kata yang bisa dikecualikan.
5. Stemming, merupakan tahapan merubah suatu kata kembali membentuk kata dasarnya.

2.5 Term Weighting

Merupakan tahapan pemberian nilai bobot terhadap kata (term) dalam sebuah dokumen, pembobotan yang digunakan dalam penelitian ini yakni TF-IDF. TF-IDF adalah metrik statistik untuk menentukan pentingnya kata atau kumpulan kata dalam dokumen, selanjutnya dapat diimplementasikan menggunakan algoritma pembelajaran mesin yang digunakan.

2.6 TF-IDF

Pembobotan yang paling populer digunakan adalah TF-IDF, pendekatan ini berasal dari TF yang merupakan istilah analisis berdasarkan frekuensi dari sebuah teks kemudian IDF adalah penentuan istilah yang berada di sebuah ruang dokumen atau spesifisitas dari dokumen tersebut, istilah pengukuran pertama kali diusulkan oleh Sparck Jones dan kemudian sekarang dikenal sebagai IDF. Alasannya adalah bahwa istilah dengan jumlah frekuensi dokumen (DF) yang tinggi mungkin bukan pembeda yang baik, dan itu harus diberikan bobot yang lebih sedikit. Oleh karena itu, IDF berarti bahwa istilah yang muncul di banyak dokumen bukanlah pembeda dokumen yang baik, fungsi IDF memberikan skor terendah untuk istilah-istilah yang muncul di beberapa dokumen dalam ruang dokumen [13].

Berikut merupakan persamaan umum TF-IDF :

$$W_{TF.IDF}(t_i, d_j) = tf_{(t_i, d_j)} \times \left(\log \frac{D}{d(t_i)}\right) \tag{1}$$

Keterangan :

W : Weight (bobot kata)

$tf_{(t_i, d_j)}$: Term Frekuensi (Frekuensi kata t pada dokumen d)

D : Jumlah seluruh dokumen

$d(t_i)$: Jumlah dokumen yang mengandung kata t

2.7 Pengembangan Model Machine Learning

Proses pembelajaran atau training model klasifikasi menggunakan metode Logistic Regression untuk mendapatkan model terbaik yang nantinya dilakukan proses untuk validasi model, setelah model terbaik didapatkan maka model tersebut disimpan, selanjutnya dijadikan sebagai model proses simulasi dan pengujian akhir dalam penelitian ini.

2.8 Logistic Regression

Analisis Regresi Logistik merupakan suatu metode analisis yang digunakan untuk mendeskripsikan hubungan antara variabel terikat yang memiliki dua kategori atau lebih dengan satu atau lebih variabel bebas, berskala kategori atau kontinu [4]. Regresi logistik bertindak sebagai model analisis prediktif, Regresi logistik diterapkan untuk menggambarkan data dan menjelaskan hubungan antara variabel biner dependen dan satu atau lebih variabel independen pada tingkat nominal, orde, interval, atau rasio [14]. Jika variabel terikatnya biner, maka regresi logistik menggunakan analisis regresi adalah suatu yang tepat, namun pada studi kasus yang terjadi variabel terikatnya memiliki tiga kelas yakni Positif, Netral dan Negatif yang artinya multiclass, maka dapat menggunakan metode Binary Classifiers for Multi-Class Classification, yang salah satu tekniknya adalah menggunakan teknik One-vs-Rest (OvR) adalah cara heuristik untuk menggunakan algoritma klasifikasi biner untuk klasifikasi multiclass. Dataset multiclass dapat dibagi menjadi beberapa masalah klasifikasi biner. Pengklasifikasi biner kemudian dilatih untuk setiap masalah klasifikasi biner dan prediksi dibuat menggunakan model yang paling handal. Berikut persamaan Logistic Regression:

Simple Linear Regression :

$$y = \alpha + \beta x \tag{2}$$

$$g(x) = \alpha + \beta x$$

Multiple Linear Regression :

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \tag{3}$$

$$g(X) = \alpha + \beta X$$

Logistic Regression :

$$g(X) = \text{sigmoid}(\alpha + \beta X) \tag{4}$$

$$sigmoid(x) = \frac{1}{1 + \exp -x}$$

Keterangan :

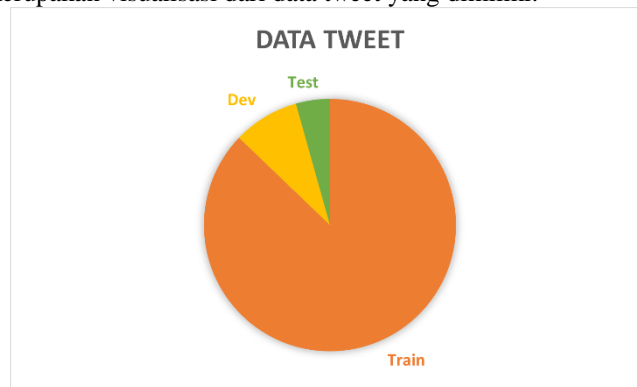
- Y : Variabel response atau variabel akibat (dependent)
- α : Konstanta
- β : Koefisien regresi (kemiringan); besaran response yang ditimbulkan oleh predictor.
- X : Variabel Predictor atau Variabel Faktor Penyebab (Independent)

2.9 Implementasi dan Pengujian

Pada bagian ini, langkah implementasi akan dilakukan. Model penelitian ini dibuat berdasarkan temuan analisis. Google Collaboratory digunakan dalam desain untuk menerapkan temuan analisis ke format baris kode pemrograman Python. Pengujian adalah tahap selanjutnya, langkah pengujian ini diperlukan sebagai acuan untuk memastikan bahwa penelitian yang dilakukan sesuai dengan harapan dan tujuan, pengujian akan dilakukan menggunakan confusion matrix.

3. HASIL DAN PEMBAHASAN

Proses yang dilakukan adalah melakukan pengolahan terhadap data sentimen, data sentimen terkumpul sebanyak 9178 data tweet yang kemudian data dikelompokkan masing-masing menjadi data train, dev serta test, pembagiannya adalah sebagai berikut data train 8000 tweet, dev 778 tweet dan test 400 tweet. Data dev digunakan untuk menentukan model terbaik yang dapat dihasilkan untuk nanti diujikan kepada data test. Pada proses klasifikasi akan digunakan algoritma Logistic Regression dengan pembobotan menggunakan tf-idf. Kemudian dari hasil klasifikasi tersebut akan dilakukan evaluasi analisis kinerja sentimen dengan menggunakan pengujian yang umumnya dilakukan yakni confusion matrix untuk mengetahui tingkat akurasi kinerja sistem yang dibangun. Berikut merupakan visualisasi dari data tweet yang dimiliki:



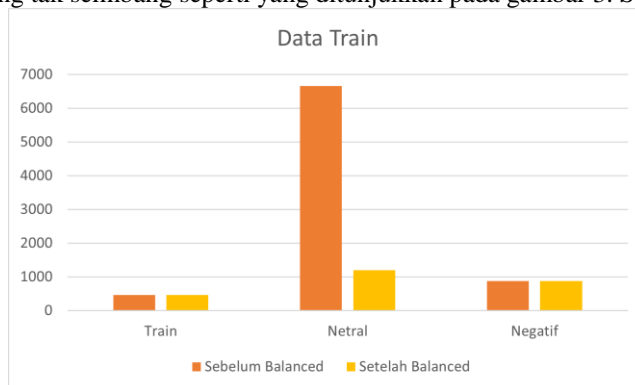
Gambar 2. Visualisasi Data

Berikut proses yang dilakukan adalah melakukan pemrosesan teks, proses yang dilakukan menggunakan kombinasi terbaik yang didapatkan dari penelitian [15] yakni kombinasi feature selection dari kombinasi case folding, remove single character serta remove punctuation. Setelah melewati proses processing maka langkah selanjutnya adalah melakukan pembobotan terhadap data, menggunakan pembobotan tf-idf, kemudian dilakukan pengklasifikasian pada tabel 2. menggunakan algoritma Logistic Regression, berikut merupakan hasil dari proses klasifikasi:

Tabel 2. Pengujian Confusion Matrix

Pengujian Terhadap	Akurasi (%)	Precision (%)	Recall (%)	F1-score (%)
Dev	82	59	40	42

Jika melihat tabel eksperimen di atas akurasi terhadap data dev cukup tinggi yakni 82% pada akurasi, tapi bukan itu saja menjadi konsentrasi, angka yang ditunjukkan oleh pengujian F1-Score kurang bagus di data dev sekitaran 42%, berikutnya dalam membangun sebuah model harus memperhatikan nilai F1-Score nya, karena distribusi data sentimen pada data train memiliki distribusi klasifikasi yang tak seimbang seperti yang ditunjukkan pada gambar 3. Sebelum dilakukan balanced.



Gambar 3. Visualisasi Data Train

Selanjutnya dalam penelitian ini akan melakukan pendekatan yang dapat meningkatkan nilai akurasi F1-Score, cara yang pertama adalah menstabilkan distribusi dari data train, maka akan dilakukan Slicing data yakni untuk mengambil sebagian dataset dari sebuah data frame, dan menyeimbangkan dengan rasio professional kelas positif, negatif dan netral antar kelasnya penyeimbangan mengacu pada penelitian [16] sehingga terbentuk data train netral yang telah diseimbangkan seperti terlihat pada gambar 3. data setelah balanced, dalam penelitian ini data train kelas netral yang diambil adalah data dengan jumlah kata dalam tweet lebih dari delapan kata sehingga dapat mereduksi tweet yang kurang informatif. Selain itu juga dilakukan pengoptimalan Tuning Hyperparameters dari algoritma Logistic Regression itu sendiri dengan mencoba parameter terbaik, parameter yang akan di uji adalah sebagai berikut:

$$\begin{aligned}
 C &= \{0.01, 0.05, 0.25, 0.5, 0.75, 1\} \\
 Penalty &= \{ 'l1', 'l2', 'elasticnet', 'none' \} \\
 Solver &= \{ 'newton - cg', 'lbfgs', 'liblinear', 'sag', 'saga' \}
 \end{aligned}
 \tag{5}$$

Pengujian dilakukan menggunakan bantuan GridSearchCV yakni berupa modul dari scikit-learn untuk memvalidasi model dan hyperparameter menggunakan teknik Cross Validation [17], menggunakan 3 Cross Validation menghasilkan 120 kandidat dan 360 fits sehingga didapatkan 5 hyperparameter terbaik pada tabel 3. yakni:

Tabel 3. Hasil GridSearchCV Hyperparameter

Id kombinasi	solver	penalty	Param_C	mean_test_score	rank_test_score
105	newton_cg	l2	1	0.631853982	1
106	lbfgs	l2	1	0.631853982	1
48	sag	l2	0.25	0.631589851	3
69	saga	l2	0.5	0.630297385	4
47	liblinear	l2	0.25	0.630184555	5

Langkah selanjutnya parameter yang dihasilkan kemudian diterapkan pada model dan dilakukan pengujian terhadap data dev. Berikut merupakan hasil dari pengoptimalan Tuning Hyperparameters:

Tabel 4. Pengujian eksperimen terhadap Data Dev

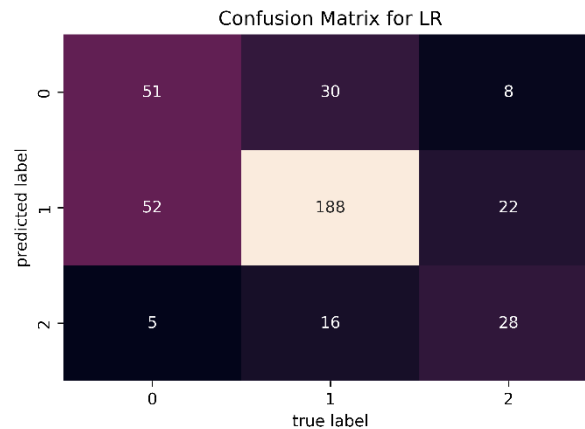
Id kombinasi	Akurasi (%)	Precision (%)	Recall (%)	F1-score (%)
105	77	56	61	56.91
106	77	56	61	56.91
48	78	55	61	56.68
69	77	56	50	56.62
47	77	55	61	56.83

Berdasarkan tabel 4. disimpulkan bahwa id 105 merupakan sampel eksperimen terbaik dalam penelitian ini dengan solver: newton-cg nilai C: 1 serta penalty: l2 terbukti dalam pengujian terhadap data dev menghasilkan f1-score sebesar 0.5683166021884959 atau 57%, id kombinasi 105 diambil karna lebih unggul dari rata-rata fit time lebih cepat dari pada kombinasi 106, model terbaik ini akan diterapkan pada data test untuk memvalidasi dari model yang dihasilkan. Berikut merupakan pengujian terhadap data test:

Tabel 5. Pengujian terhadap Data Test

Dataset	Akurasi (%)	Precision (%)	Recall (%)	F1-score (%)
Test	67	62	59	60

Pengujian terhadap data test menghasilkan akurasi 67% diikuti precision 62%, recall 59% serta nilai f1-score yang cukup baik yakni 60% ini membuktikan bahwa model yang dibangun cukup handal dalam melakukan klasifikasi seperti pada tabel 5. terbukti nilai f1-score cukup berimbang dan berikut merupakan visualisasi pengujian confusion matrix:



Gambar 4. Confusion Matrix Data Test

Berdasarkan dari eksperimen diatas metode Logistic Regression cukup handal dalam membangun model atau bisa bersaing dengan penelitian sebelumnya yakni penelitian Naïve Bayes [15], LSTM [16], SVM [18] sebagaimana berikut:

Tabel 6. Perbandingan Metode

Metode	Akurasi	Precision	Recall	F1-score
Naïve Bayes [15]	0.61	0.58	0.60	0.57
SVM [18]	0.65	0.61	0.54	0.56
LSTM [16]	0.66	0.75	0.53	0.54
LR (Metode ini)	0.67	0.62	0.59	0.60

Disimpulkan dari tabel 6. bahwa metode ini cukup baik dalam melakukan pemodelan berdasarkan beberapa penelitian sebelumnya dimana akurasi yang dihasilkan lebih baik dari pada metode sebelumnya yakni 67% serta nilai f1-score juga menghasilkan nilai terbaik yakni 60% selain bisa bersaing dengan metode konvensional lainnya, terbukti LR juga dapat bersaing dengan metode modern atau terbaru seperti LSTM.

4. KESIMPULAN

Penelitian sentimen terhadap vaksinasi Covid Twitter telah berhasil dibangun dengan menerapkan algoritma Logistic Regression, dimana akurasi tertinggi didapatkan di angka 82% data dev artinya algoritma ini cukup baik dalam melakukan klasifikasi, tapi tidak diimbangi dengan F1-Score yang cukup baik yakni 42% terhadap data dev, hal ini terjadi karna terjadinya tidak keseimbangan data atau imbalance pada data train atau data pembelajaran pada saat membangun modelnya, maka dilakukan proses menyeimbangkan data serta dalam penelitian ini dan juga melakukan Tuning Hyperparameter yang bertujuan untuk meningkatkan tingkat akurasi dari F1-Score dari proses sebelumnya, setelah dilakukan penyeimbangan data dan menerapkan parameter terbaik yang diperoleh kedalam model, metode Logistic Regression menghasilkan nilai akurasi yang cukup baik dari pada penelitian sebelumnya dimana F1-score 60% dan akurasi 67% terhadap data test hasil ini membuktikan bahwa model yang dibangun cukup handal dalam melakukan klasifikasi, terbukti nilai f1-score dan akurasi cukup berimbang dan lebih baik dari Naïve Bayes, SVM dan LSTM.

DAFTAR PUSTAKA

- [1] I. P. Sari and S. Sriwidodo, "Perkembangan Teknologi Terkini dalam Mempercepat Produksi Vaksin COVID-19," *Maj. Farmasetika*, vol. 5, no. 5, p. 204, 2020, doi: 10.24198/mfarmasetika.v5i5.28082.
- [2] Kementerian Kesehatan Republik Indonesia, "Peraturan Menteri Kesehatan Republik Indonesia Nomor 10 Tahun 2021," Jakarta, 2021. [Online]. Available: <https://peraturan.go.id/peraturan/view.html?id=65bb78cf233585b52b836b1315fd6a75>.
- [3] Kementerian Kesehatan Republik Indonesia, "Survei Penerimaan Vaksin COVID-19 di Indonesia," Jakarta, 2020. [Online]. Available: <https://www.unicef.org/indonesia/id/coronavirus/laporan/survei-penerimaan-vaksin-covid-19-di-indonesia>.
- [4] A. Syahadati, N. C. Lengkong, O. Safitri, S. Machsus, Y. R. Putra, and R. Nooraeni, "Analisis Sentimen Penerapan PSBB di DKI Jakarta dan Dampaknya Terhadap Pergerakan IHSG," vol. 15, no. 1, pp. 20–25, 2021, doi: 10.33365/jti.v15i1.866.
- [5] Y. A. Rahman, "Vaksinasi Massal Covid-19 sebagai Sebuah Upaya Masyarakat dalam Melaksanakan Kepatuhan Hukum (Obedience Law)," *Khazanah Huk.*, vol. 3, no. 2, pp. 80–86, 2021, doi: 10.15575/kh.v3i2.11520.
- [6] N. P. Astuti, E. G. Z. Nugroho, J. C. Lattu, I. R. Potempu, and D. A. Swandana, "Persepsi Masyarakat terhadap Penerimaan Vaksinasi Covid-19: Literature Review," *J. Keperawatan*, vol. 13, no. 3, pp. 569–580, 2021, doi: 10.32583/keperawatan.v13i3.1363.
- [7] B. Liu and L. Zhang, *Mining text data*, vol. 9781461432. 2013.
- [8] H. Rianto and R. S. Wahono, "Resampling Logistic Regression untuk Penanganan Ketidakseimbangan Class pada Prediksi Cacat Software," *IlmuKomputer.com J. Softw. Eng.*, vol. 1, no. 1, pp. 46–53, 2015.
- [9] H. Hasanli and S. Rustamov, "Sentiment Analysis of Azerbaijani tweets Using Logistic Regression , Naive Bayes and SVM," in *2019 IEEE 13th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 1–7.
- [10] E. D. Wardhani, S. K. Areka, A. W. Nugroho, A. R. Zakaria, A. D. Prakasa, and R. Nooraeni, "Sentiment Analysis Using Twitter Data Regarding BPJS Cost Increase and Its Effect on Health Sector Stock Prices," *Indones. J. Artif. Intell. Data Min.*, vol. 3, no. 1, p. 1, 2020, doi: 10.24014/ijaidm.v3i1.8245.
- [11] S. R. Dewi, N. Salam, and D. S. Susanti, "Klasifikasi Pemilihan Program Studi di Fakultas MIPA Universitas Lambung Mangkurat Menggunakan Regresi Logistik Multinomial," *J. Mat. Murni dan Terap. "epsilon"*, vol. 12, no. 2, pp. 19–29, 2018.
- [12] A. G. Gani *et al.*, "Klasifikasi jamur menggunakan algoritma naïve bayes," *J. CoSciTech (Computer Sci. Inf. Technol.)*, vol. 3, no. 2, pp. 28–33, 2022, doi: 10.37859/coscitech.v3i1.3685.
- [13] F. Ren and M. G. Sohrab, "Class-indexing-based term weighting for automatic text classification," *Inf. Sci. (Ny.)*, vol. 236, pp. 109–125, 2013, doi: 10.1016/j.ins.2013.02.029.
- [14] T. Ciu and R. S. Oetama, "Logistic Regression Prediction Model for Cardiovascular Disease," *IJNMT (International J. New Media Technol.)*, vol. 7, no. 1, pp. 33–38, 2020, doi: 10.31937/ijnmt.v7i1.1340.
- [15] P. Yohana, "Analisis Sentimen Masyarakat Terhadap Kebijakan Pemerintah Indonesia Dalam Memberikan Vaksin Covid-19 Menggunakan Metode Naive Bayes Classifier," Universitas Islam Negeri Sultan Syarif Kasim Riau, 2021.
- [16] M. Ihsan, B. S. Negara, and S. Agustian, "Metode LSTM (Long short term memory) untuk Klasifikasi Sentimen Vaksin Covid-19 pada Twitter," pp. 1–13, 2022, doi: 10.31849/digitalzone.v13i1.9950.
- [17] Aditya Quantano Surbakti, Regiolina Hayami, and Januar Al Amien, "Analisa Tanggapan Terhadap Psbb Di Indonesia Dengan Algoritma Decision Tree Pada Twitter," *J. CoSciTech (Computer Sci. Inf. Technol.)*, vol. 2, no. 2, pp. 91–97, 2021, doi: 10.37859/coscitech.v2i2.2851.
- [18] M. Rizki, "Analisis Sentimen Masyarakat Terhadap Vaksin Covid-19 Menggunakan Metode Support Vector Machine Pada Media Sosial Twitter," Universitas Islam Negeri Sultan Syarif Kasim Riau, 2022.