https://jurnal.pustakagalerimandiri.co.id/index.php/pustakaai
DOI : https://doi.org/10.55382/jurnalpustakaai.v6i1.1566

JURNAL PUSTAKA AI
JURNAL PUSAT AKSES KAJIAN TEKNOLOGI ARTIFICIAL INTELLIGENCE
Pustaka Galeri Mandiri
Vol. 6 No. 1 (2026) 92 – 102
E ISSN: 2809-4069

# Performa Logistic Regression dalam Klasifikasi Sentimen Opini Publik Pemilu di India

Okky Puspa Ningrum[1], Dwi Remawati[2], Teguh Susyanto[3], Wawan Laksito Yuly Saptomo[4]
[1,3]Sistem Informasi, Fakultas Teknik, Universitas Tiga Serangkai, Surakarta
[2]Teknologi Informasi, Sekolah Vokasi, Universitas Tiga Serangkai, Surakarta
[4]Informatika, Fakultas Teknik, Universitas Tiga Serangkai, Surakarta
[1]23400014.okky@tsu.ac.id. [2]dwirema@tsu.ac.id . [3]teguh@tsu.ac.id . [4]wawanlaksito@tsu.ac.id

***Abstract***

*Public opinion about 2019 Indian General Election on social media particularly Twitter, is highly massive and dynamic. This study aims to classify these opinions into positive, negative or neutral sentiments by using the Logistic Regression algorithm. The dataset used is sourced from Kaggle. The data preprocessing methods applied include the removal missing value, tokenizing, stop-word removal and stemming which produced clean data for analysis. Text features were extracted by using TF-IDF and model performance was evaluated by using a Confusion Matrix to measure accuracy, precision, and recall. Experimental results show that the model achieved an accuracy of 84.46%. This study proves that Logistic Regression is capable of providing stable and efficient classification performance for automated public opinion monitoring on noisy social media texts.*

*Keywords: Sentiment Analysis, Twitter, Logistic Regression, TF-IDF, Classification.*

**Abstrak**

Opini publik mengenai Pemilihan Umum India 2019 di media sosial, khususnya Twitter, sangat masif dan dinamis. Penelitian ini bertujuan untuk mengklasifikasikan opini tersebut ke dalam sentimen positif, negatif, atau netral menggunakan algoritma Logistic Regression. Dataset yang digunakan bersumber dari Kaggle. Metode prapemrosesan data yang diterapkan meliputi penghapusan missing value, tokenizing, stopword removal, dan stemming, yang menghasilkan data bersih untuk dianalisis. Fitur teks diekstraksi menggunakan TF-IDF dan performa model dievaluasi menggunakan Confusion Matrix untuk mengukur tingkat akurasi, presisi, dan recall. Hasil eksperimen menunjukkan bahwa model mencapai akurasi sebesar 84,46%. Penelitian membuktikan bahwa Logistic Regression mampu memberikan performa klasifikasi yang stabil dan efisien untuk pemantauan opini publik otomatis pada teks media sosial yang sarat noise.

Kata kunci: Analisis Sentimen, Twitter, Logistic Regression, TF-IDF, Klasifikasi.



Submitted : 14-04-2026 | Reviewed : 24-04-2026 | Accepted : 30-04-2026

Jurnal Pustaka AI Vol . 6 No. 1 (2026) 92 – 102

## 1. Pendahuluan

Pemilihan Umum di India merupakan salah satu peristiwa politik terbesar di dunia, dengan ratusan juta pemilih yang aktif menyuarakan opini melalui berbagai platform digital. Media sosial seperti Twitter, Facebook, dan YouTube menjadi sarana utama masyarakat untuk mengekspresikan preferensi politik, kritik kebijakan, hingga dukungan terhadap kandidat tertentu[1], [2], [3]. Besarnya aktivitas komunikasi politik tersebut menjadikan analisis sentimen sebagai pendekatan penting untuk memahami pola opini publik secara cepat dan berbasis data[4].Analisis sentimen telah banyak digunakan dalam studi politik digital untuk mengukur persepsi masyarakat terhadap isu tertentu maupun kandidat politik[3], [5]. Berbagai metode telah diterapkan, mulai dari teknik machine learning klasik hingga pendekatan deep learning modern[6], [7].

Pada penelitian-penelitian sebelumnya, metode seperti Naïve Bayes, Support Vector Machine (SVM), dan Random Forest menunjukkan performa yang cukup baik dalam klasifikasi sentimen, terutama pada teks pendek media sosial[8], [9]. Namun, metode tersebut memiliki keterbatasan, terutama dalam menangkap makna kontekstual, menghadapi noise seperti slang, singkatan, dan struktur kalimat tidak baku yang dominan dalam percakapan politik. Di sisi lain, Logistic Regression tetap menjadi salah satu algoritma paling populer dalam analisis sentimen berbasis machine learning karena kesederhanaan, interpretabilitas yang tinggi, [5], [10] serta efisiensi komputasi pada data berdimensi tinggi dibandingkan model *deep learning* yang membutuhkan sumber daya besar. Model ini bekerja dengan baik pada representasi teks berbasis fitur seperti TF-IDF[2], [11]. Logistic Regression juga memiliki kelebihan dalam mengelola data berdimensi tinggi, menghasilkan probabilitas klasifikasi yang jelas, serta relatif lebih efisien dibandingkan model deep learning[12]

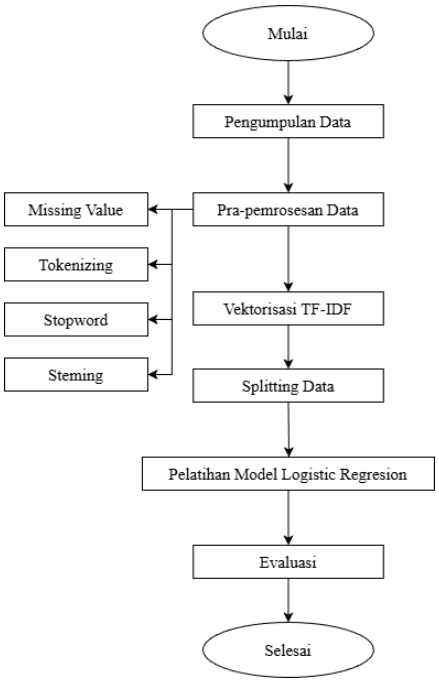
Meskipun demikian, beberapa tantangan masih ditemukan dalam penelitian terdahulu, kurangnya evaluasi komprehensif terhadap stabilitas Logistic Regression khusus pada konteks dinamika politik negara berkembang yang sangat fluktuatif seperti Pemilu India. Kebaruan dari penelitian ini terletak pada pengujian ketangguhan model melalui integrasi tahap text preprocessing yang mendalam dan pembobotan fitur TF-IDF untuk menangani karakteristik teks media sosial yang sangat dinamis. Kontribusi penelitian ini adalah memberikan kerangka kerja klasifikasi yang efisien untuk memantau opini publik secara otomatis, yang tidak hanya akurat tetapi juga memiliki kecepatan komputasi yang andal untuk dataset berskala besar.

Tujuan penelitian ini adalah untuk menganalisis pola sentimen opini publik terkait Pemilu India berdasarkan data Twitter serta mengevaluasi performa algoritma Logistic Regression melalui berbagai metrik evaluasi. Selain itu, penelitian ini bertujuan untuk mengidentifikasi kelemahan model dalam klasifikasi sentimen tertentu guna memberikan rekomendasi bagi pengembangan sistem pemantauan opini digital yang lebih stabil di masa depan. Dengan demikian, hasil penelitian ini diharapkan dapat membantu analis politik dan pemangku kebijakan dalam memahami dinamika persepsi masyarakat secara real-time.

## 2. Metode Penelitian

Bagian ini merepresentasikan langkah-langkah dan tahapan yang dilalui secara sistematis pada studi analisis sentimen Twitter ini. Proses keseluruhan dimulai dari pencarian data hingga pengujian dan evaluasi model klasifikasi. Urutan proses tersebut digambarkan secara visual melalui Gambar 1. Proses Analisis Sentimen, yang memandu alur kerja penelitian.

Submitted : 14-04-2026 | Reviewed : 24-04-2026 | Accepted : 30-04-2026




Gambar 1. Flowchart Proses Analisis Sentimen

Pada flowchart menjelaskan tahapan dari awal penelitian hingga selesai penelitian yang diuraikan sebagai berikut:

1) Mencari Dataset

Dataset yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari Kaggle (Twitter Sentiment Dataset) data ini merupakan data sekunder yang sudah terlabel secara prapabrikan (pre-labeled) ke dalam tiga kategori, yaitu positif (1.0), netral (0.0), dan negatif (-1.0). Oleh karena itu, peneliti tidak melakukan proses pelabelan manual, melainkan berfokus pada tahap prapemrosesan (preprocessing) dan pemodelan menggunakan Logistic Regression untuk menguji akurasi prediksi pada label yang tersedia. Studi diterapkan dengan bahasa pemrograman Python melalui platform Google Colab untuk memastikan efisiensi komputasi pada dataset besar. Penggunaan dataset publik Kaggle adalah praktik umum dalam riset machine learning untuk memastikan efisiensi komputasi pada dataset besar.[9], [13]

Tabel 1. Tampilan Data Mentah

| | clean_text | category |
|---|---|---|
| **0** | when modi promised “minimum government maximum... | -1.0 |
| **1** | talk all the nonsense and continue all the dra... | 0.0 |
| **2** | what did just say vote for modi welcome bjp t... | 1.0 |
| **3** | asking his supporters prefix chowkidar their n... | 1.0 |
| **4** | answer who among these the most powerful world... | 1.0 |
| **...** | ... | ... |
| **162975** | why these 456 crores paid neerav modi not reco... | -1.0 |
| **162976** | dear rss terrorist payal gawar what about modi... | -1.0 |
| **162977** | did you cover her interaction forum where she ... | 0.0 |
| **162978** | there big project came into india modi dream p... | 0.0 |
| **162979** | have you ever listen about like gurukul where ... | 1.0 |

162980 rows × 2 columns

2) Pra-pemrosesan Data

Dalam penelitian ini, terdapat beberapa alur proses krusial untuk menentukan kualitas hasil akhir. Tahapan ini merupakan proses pembersihan data guna mengkonversi teks menjadi representasi yang lebih terstruktur sehingga proses klasifikasi dapat berjalan secara efektif dan efisien, serta untuk memastikan hasil akurat dan berkualitas. Hal ini sejalan dengan penelitian [3] yang menekankan bahwa kualitas data pra-pemrosesan berbanding lurus dengan tingkat akurasi model. Berikut tahapan preprocessing data yang dilakukan:

a. Missing Value: Pada tahap ini, data yang memiliki nilai kosong (null atau NaN) pada kolom clean_text dan category dihapus agar tidak mempengaruhi hasil analisis.

Tabel 2. Tampilan data yang mengandung missing value

| | clean_text | category |
|---|---|---|
| **148** | NaN | 0.0 |
| **130448** | the foundation stone northeast gas grid inaugu... | NaN |
| **155642** | dear terrorists you can run but you cant hide ... | NaN |
| **155698** | offense the best defence with mission shakti m... | NaN |
| **155770** | have always heard politicians backing out thei... | NaN |
| **158693** | modi government plans felicitate the faceless ... | NaN |
| **158694** | NaN | -1.0 |
| **159442** | chidambaram gives praises modinomics | NaN |
| **159443** | NaN | 0.0 |
| **160559** | the reason why modi contested from seats 2014 ... | NaN |
| **160560** | NaN | 1.0 |

b. Tokenizing: proses pemecah teks menjadi bagian-bagian kecil berupa kata-kata yang disebut token. Token biasanya berupa kata, frasa, kalimat, atau simbol, bergantung pada konteks serta tujuan analisis.
c. Stopword: proses menghilangkan kata-kata umum yang sering muncul tetapi tidak memiliki bobot penting bagi analisis sentimen, seperti is, are, the, dan of.
d. Stemming: proses mengembalikan kata ke bentuk asalnya (root word) agar variasi kata yang memiliki makna identik diperlakukan sebagai satu bentuk.

Tabel 3. Hasil Transformasi Data Teks (Tokenizing, Stopword, Stemming)

| | clean_text | category | stemmed_content |
|---|---|---|---|
| **0** | when modi promised “minimum government maximum... | -1.0 | modi promis minimum govern maximum govern expe... |
| **1** | talk all the nonsense and continue all the dra... | 0.0 | talk nonsens continu drama vote modi |
| **2** | what did just say vote for modi welcome bjp t... | 1.0 | say vote modi welcom bjp told rahul main campa... |
| **3** | asking his supporters prefix chowkidar their n... | 1.0 | ask support prefix chowkidar name modi great s... |
| **4** | answer who among these the most powerful world... | 1.0 | answer among power world leader today trump pu... |
| **...** | ... | ... | ... |

Submitted : 14-04-2026 | Reviewed : 24-04-2026 | Accepted : 30-04-2026

| | clean_text | category | stemmed_content |
|---|---|---|---|
| **162975** | why these 456 crores paid neerav modi not reco... | -1.0 | crore paid neerav modi recov congress leader h... |
| **162976** | dear rss terrorist payal gawar what about modi... | -1.0 | dear rss terrorist payal gawar modi kill plu m... |
| **162977** | did you cover her interaction forum where she ... | 0.0 | cover interact forum left |
| **162978** | there big project came into india modi dream p... | 0.0 | big project came india modi dream project happ... |
| **162979** | have you ever listen about like gurukul where ... | 1.0 | ever listen like gurukul disciplin maintain ev... |

162969 rows × 3 columns

3) Vektorizer TF-IDF
Data teks yang telah bersih kemudian dikonversi menjadi representasi numerik menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF digunakan untuk memberikan bobot pada setiap kata berdasarkan frekuensi kemunculannya. Teknik ini terbukti sangat efektif dalam menonjolkan fitur kata kunci yang mencerminkan sentimen atau kondisi tertentu dalam teks[2], [11].

4) Splitting Data
Untuk menguji validitas model, dataset yang telah diproses dibagi menjadi dua bagian, yaitu data latih (training set) dan data uji (testing set). Penelitian ini menerapkan rasio pembagian 80:20, di mana 80% data digunakan untuk melatih algoritma Logistic Regression dalam mengenali pola sentimen, sementara 20% sisanya digunakan sebagai instrumen evaluasi objektif untuk mengukur performa model pada data yang belum pernah dilihat sebelumnya[8], [9].

5) Klasifikasi Model Logistic Regression
Proses klasifikasi menggunakan algoritma Logistic Regression. Algoritma ini dipilih karena stabilitas dalam menangani data teks berskala besar dengan waktu komputasi yang efisien. Pemilihan model yang tepat sangat penting dalam membangun sistem prediksi yang andal, baik untuk analisis sentimen maupun prediksi data terstruktur lainnya [14] Berikut rumusnya:
   a. Sigmoid

$$Sigmoid = P(y = 1|x) = \sigma(z) = \frac{1}{1+e^{-z}} \quad (1)$$

   b. Logit

$$z = \beta_0 + \beta_{1x1} + \beta_{2x2} + \cdots + \beta_n x_n \quad (2)$$

   c. Log-Odds

$$Odds = In\left(\frac{P}{1-P}\right) = \beta_0 + \beta_{1x1} + \cdots + \beta_{n\ n}\ _{x} \quad (3)$$

6) Evaluasi Model
Evaluasi dilakukan menggunakan Confusion Matrix untuk mengukur performa akurasi, presisi, dan recall guna memastikan model mampu mengenali setiap kategori sentimen dengan baik.
Berikut adalah rumusnya[9]:

   a. Acuracy

$$Accuracy = \frac{TP_{\underline{Neg}} \pm TP_{\underline{Net}} \pm TP_{\underline{pos}}}{Total\ Sampel} \quad (4)$$

b. Precision

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

c. Recall

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

d. F1-Score

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

e. Macro Avg

$$Macro\ Avg = \frac{\sum_{i=1}^{n} Metric_i}{n} \quad (8)$$

f. Weighted Avg

$$Weighted\ Avg = \frac{\sum(Score_i \times Support_i)}{Total\ Support} \quad (9)$$

## 3. Hasil dan Pembahasan

Bagian ini memaparkan hasil eksperimen analisis sentimen menggunakan algoritma Logistic Regression pada data Twitter Pemilu India 2019. Dataset mentah yang digunakan semula berjumlah 162.980 baris. Namun, setelah melalui tahapan preprocessing, ditemukan 11 baris data yang mengandung missing value sehingga baris tersebut dihapus menggunakan fungsi df.dropna(). Hal ini menyebabkan jumlah data final yang digunakan dalam analisis menjadi 162.969 baris data.

### 3.1. Distribusi Sentimen

Berdasarkan hasil setelah pemrosesan data, dilakukan penghitungan jumlah label pada kolom kategori menggunakan fungsi value.counts().

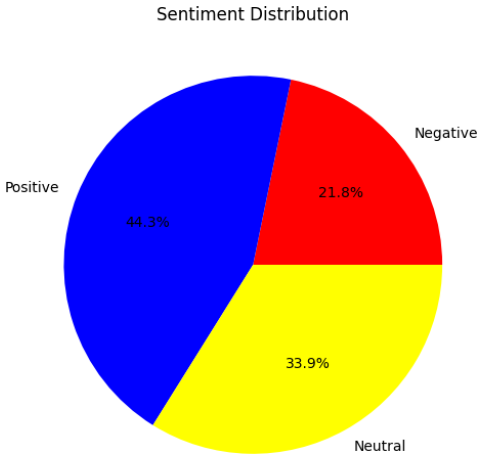
Tabel 4. Distribusi jumlah label per category

| | **count** |
|---|---|
| **category** | |
| **1.0** | 72249 |
| **0.0** | 55211 |
| **-1.0** | 35509 |

**dtype:** int64

Distribusi data menunjukan dominasi sentimen positif (1.0) sebanyak 72.294 data, diikuti oleh sentimen netral (0.0) sebanyak 55.211 data, dan sentimen negatif (-1.0) sebanyak 35.509 data. Untuk mempermudah pemahaman visual, hasil ini dikonversikan ke dalam bentuk persentase melalui pie chart.

Submitted : 14-04-2026 | Reviewed : 24-04-2026 | Accepted : 30-04-2026

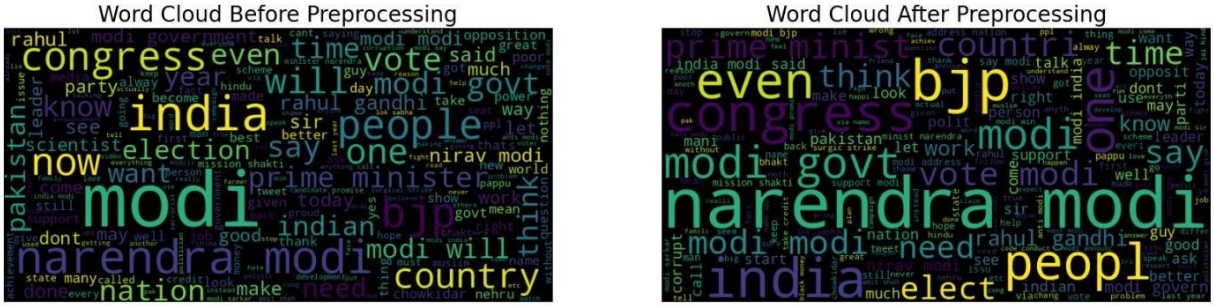



Gambar 2. Distribusi jumlah dengan persen

Visualisasi ini menunjukkan bahwa mayoritas opini publik selama Pemilu India 2019 cenderung memberikan dukungan atau respons positif. Keidakseimbangan jumlah data antar kelas (imbalance data) merupakan karakteristik umum dalam dataset media sosial berskala besar, yang juga ditemukan dalam penelitian analisis sentimen Twitter lainnya [3], [8].

3.2. Visualisasi WordCloud

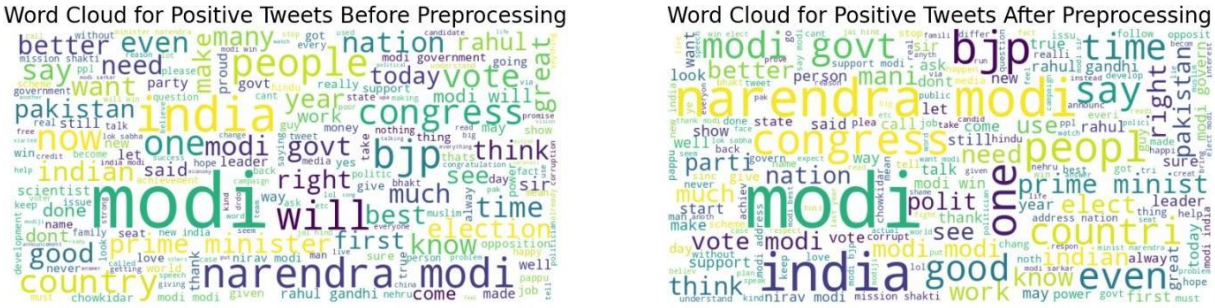
Analisis kata yang paling sering muncul dilakukan menggunakan WordCloud[15]. Visualisasi ini membantu mengidentifikasi kata kunci dominan sebelum dan sesudah tahap processing




Gambar 3. Visual WordCloud Before After Preprocessing

Penjelasan: Melalui Wordcloud, terlihat kata-kata seperti modi, india, even, bjb, congress memiliki ukuran yang paling besar. Hal ini menunjukkan bahwa fokus utama opini publik dalam dataset berkaitan erat dengan momentum politik tersebut

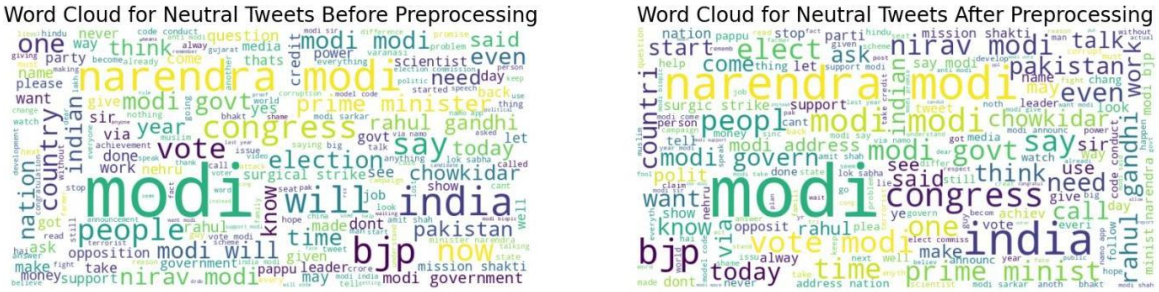
1. Sentimen Positif




Gambar 4. Visual Before After Positif

Penjelasan: Setelah preprocessing, WordCloud positif menjadi sangat terfokus pada modi dan narendra, menunjukkan bahwa sentimen dukungan terkonsentrasi pada tokoh politik utama dan partai bjp.

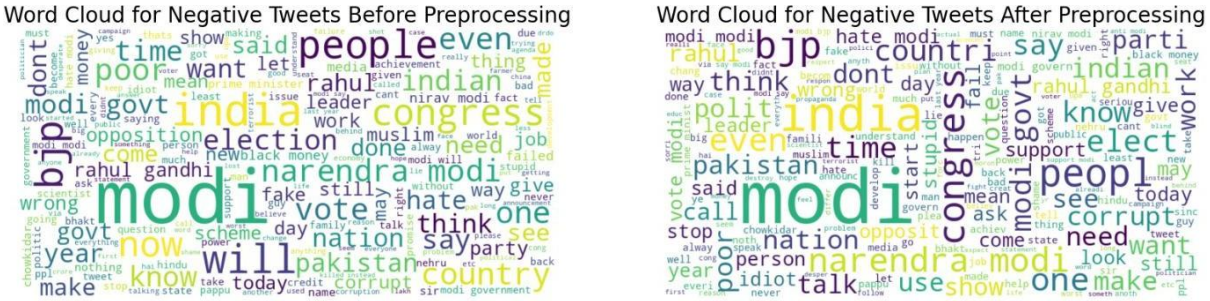
2. Sentimen Netral




Gambar 5. Visual Before After Netral

Penjelasan: Setelah preprocessing, WordCloud netral menjadi kata-kata informatif yang cenderung melaporkan fakta atau berita tanpa memberikan mautan emosional yang kuat.

3. Sentimen Negatif

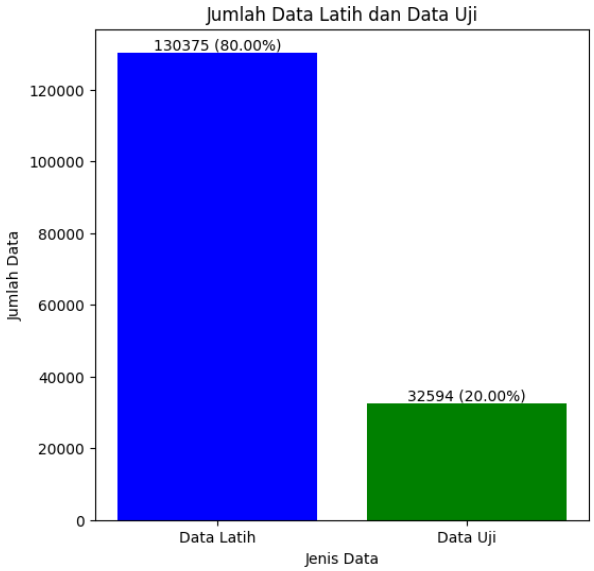



Gambar 6. Visual Before After Negatif

Penjelasan: Setelah preprocessing, WordCloud negatif menonjolkan kata-kata yang mengandung kritik, ketidakpuasan, atau isu-isu kontroversial yang muncul selama periode pemilu

3.3. Pembagian Data (Data Splitting)

Proporsi Dataset dibagi menjadi data latih dan data uji dengan rasio **80:20**.




Gambar 7. Distribusi Pembagian Data Latih dan Data Uji

Berdasarkan gambar 7, distribusi data terlihat proporsional dimana data latih memiliki jumlah sampel yang jauh lebih besar. Penggunaan rasio 80:20 ini didasarkan pada standar eksperimen machine learning untuk dataset besar agar model terhindar dari overfitting dan tetap stabil saat melakukan prediksi [8], [9].

3.4 Penerapan TF-IDF

Pengunaan TF-IDF sangat efektif dalam menyeleksi kata-kata kunci yang memiliki makna kuat terhadap sentimen tertentu dan mengabaikan kata-kata yang terlalu umum. Hal ini sejalan dengan penelitian [2], [11]yang menyatakan bahwa ekstraksi fitur dengan TF-IDF mampu meningkatkan performa klasifikasi teks secara signifikan dengan memberikan bobot yang tepat pada setiap fitur kata. Selain itu, [8] juga menekankan bahwa TF-IDF merupakan standar yang andal dalam analisis sentimen Twitter untuk menangani data teks yang beragam. Hasil TF-IDF seperti pada gambar 8:

```
=== TF-IDF 10 baris pertama ===
   aap  abl      abus  accept  account  achiev  act  action  actual  address  \
0  0.0  0.0  0.000000     0.0      0.0     0.0  0.0     0.0     0.0      0.0
1  0.0  0.0  0.000000     0.0      0.0     0.0  0.0     0.0     0.0      0.0
2  0.0  0.0  0.000000     0.0      0.0     0.0  0.0     0.0     0.0      0.0
3  0.0  0.0  0.339993     0.0      0.0     0.0  0.0     0.0     0.0      0.0
4  0.0  0.0  0.000000     0.0      0.0     0.0  0.0     0.0     0.0      0.0
5  0.0  0.0  0.000000     0.0      0.0     0.0  0.0     0.0     0.0      0.0
6  0.0  0.0  0.000000     0.0      0.0     0.0  0.0     0.0     0.0      0.0
7  0.0  0.0  0.000000     0.0      0.0     0.0  0.0     0.0     0.0      0.0
8  0.0  0.0  0.000000     0.0      0.0     0.0  0.0     0.0     0.0      0.0
9  0.0  0.0  0.000000     0.0      0.0     0.0  0.0     0.0     0.0      0.0

   ...     world  worri  would  write  wrong   ye      year  yet  youth   yr
0  ...  0.000000    0.0    0.0    0.0    0.0  0.0  0.215285  0.0    0.0  0.0
1  ...  0.000000    0.0    0.0    0.0    0.0  0.0  0.000000  0.0    0.0  0.0
2  ...  0.000000    0.0    0.0    0.0    0.0  0.0  0.000000  0.0    0.0  0.0
3  ...  0.000000    0.0    0.0    0.0    0.0  0.0  0.000000  0.0    0.0  0.0
4  ...  0.404477    0.0    0.0    0.0    0.0  0.0  0.000000  0.0    0.0  0.0
5  ...  0.000000    0.0    0.0    0.0    0.0  0.0  0.000000  0.0    0.0  0.0
6  ...  0.000000    0.0    0.0    0.0    0.0  0.0  0.000000  0.0    0.0  0.0
7  ...  0.000000    0.0    0.0    0.0    0.0  0.0  0.000000  0.0    0.0  0.0
8  ...  0.000000    0.0    0.0    0.0    0.0  0.0  0.000000  0.0    0.0  0.0
9  ...  0.000000    0.0    0.0    0.0    0.0  0.0  0.000000  0.0    0.0  0.0

[10 rows x 500 columns]
```

Gambar 8. Hasil TF-IDF

3.5 Hasil Klasifikasi Logistic Regression dan Evaluasi

Hasil implementasi Logistic Regression menunjukkan performa yang sangat impresif. Berdasarkan pengujian pada data uji, model berhasil mencapai tingkat akurasi seperti pada gambar 9

```
Accuracy: 0.84
Accuracy: 84.46%
```

Gambar 9. Hasil Accuracy

Rincian performa model disajikan dalam classification report pada gambar 10

```
Classification Report Logistic Regression:
              precision    recall  f1-score   support

        -1.0       0.82      0.71      0.76      7102
         0.0       0.82      0.91      0.86     11042
         1.0       0.88      0.86      0.87     14450

    accuracy                           0.84     32594
   macro avg       0.84      0.83      0.83     32594
weighted avg       0.85      0.84      0.84     32594
```

Gambar 10. Performa Classification Report

Yang kemudian dikonversi kedalam persentase sebagai berikut:

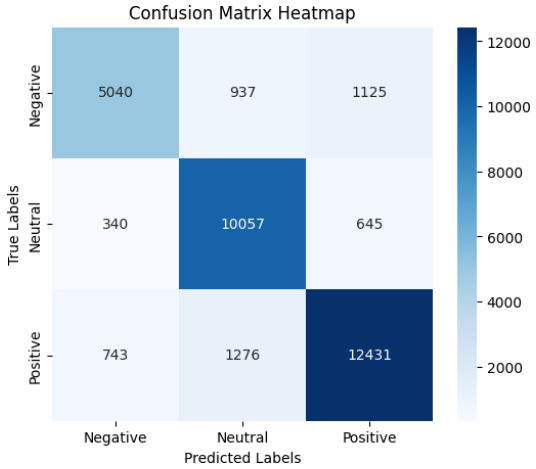
Tabel 5. Konversi hasil persentase klasifikasi

| | **precision** | **recall** | **f1-score** | **support** |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| **-1.0** | 82.31% | 70.97% | 76.22% | 7102.000000 |
| **0.0** | 81.96% | 91.08% | 86.28% | 11042.000000 |
| **1.0** | 87.54% | 86.03% | 86.78% | 14450.000000 |
| **accuracy** | 84.46% | 84.46% | 84.46% | 0.844573 |
| **macro avg** | 83.94% | 82.69% | 83.09% | 32594.000000 |
| **weighted avg** | 84.51% | 84.46% | 84.31% | 32594.000000 |

Confusion Matrix

Evaluasi akhir dilakukan dengan memanfaatkan Confusion Matrix guna meninjau sebaran prediksi benar dan salah pada tiap label (Positif, Netral, Negatif)




Gambar 11. Visual Confusion Matrix

Dari hasil visual, Confusion Matrix sangat krusial untuk membuktikan bahwa akurasi yang diperoleh bukan sekedar angka acak, melainkan hasil dari kemampuan model dalam membedakan karakteristik tiap kelas secara tepat. Sebagaimana dijelaskan oleh [8], visualisasi membantu mengidentifikasi kelas nama yang memerlukan perhatian lebih, yang dalam hal ini adalah kelas negatif karena memiliki jumlah sampel (support) yang paling sedikit dibandingkan kelas lainnya.

## 4. Kesimpulan

Berdasarkan hasil penelitian, dapat disimpulkan bahwa algoritma Logistic Regression efektif untuk klasifikasi sentimen politik berskala besar dengan tingkat akurasi 84,46%. Model ini sangat andal dalam mengenali sentimen positif dan netral, sehingga layak digunakan oleh para pemangku kepentingan sebagai alat otomatisasi pelacak persepsi masyarakat secara real-time. Namun, kelemahan pada identifikasi sentimen negatif menunjukkan perlunya peningkatan pada teknik penyeimbangan data (oversampling) atau penggunaan fitur n-gram untuk menangkap konteks negasi yang lebih kompleks pada penelitian selanjutnya.

## Ucapan Terimakasih

Penulis mengucapkan terimakasih kepada dosen pengampu mata kuliah Data Science atas fasilitas akademik yang ditawarkan serta pihak yang terlibat pada penelitian ini.

## Daftar Rujukan

[1] A. Khare, A. Gangwar, S. Singh, and S. Prakash, “Sentiment Analysis and Sarcasm Detection in Indian General Election Tweets,” *Research Advances in Intelligent Computing*, pp. 253–268, 2023, doi: 10.1201/9781003320340-20.

[2] D. Remawati, H. Wijayanto, Y. R. Wahyu Utami, and B. D. Raharja, "Pengelompokkan Film Trending di Youtube Menggunakan TF-IDF dan K-Means Clastering," *Jurnal Sistem Informasi Triguna Dharma (JURSI TGD)*, vol. 4, no. 1, pp. 65–74, 2025, doi: 10.53513/jursi.v4i1.10614.

[3] N. L. P. C. Savitri, R. A. Rahman, R. Venyutzky, and N. A. Rakhmawati, "Analisis Klasifikasi Sentimen Terhadap Sekolah Daring pada Twitter Menggunakan Supervised Machine Learning," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 7, no. 1, Apr. 2021, doi: 10.28932/jutisi.v7i1.3216.

[4] R. A. Nugroho and A. Mufliq, "Penerapan Ant Colony Optimization dengan Sentiment-Based Weighting untuk Rekomendasi Rute Wisata," *Rizky Aditya Nugroho*, vol. 1, no. 2, pp. 364–371, 2025, [Online]. Available: https://doi.org/10.55382/jurnalpustakaai.v5i2.1197

[5] A. R. Hidayati, A. S. Fitrani, and M. A. Rosid, "Analisa Sentimen Pemilu 2019 Pada Judul Berita Online Menggunakan Metode Logistic Regression [Sentiment Analysis of the 2019 Election in Online News Headlines Using the Logistic Regression Method]," *KESATRIA Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, vol. 4, no. 2, pp. 1–6, 2023, [Online]. Available: https://tunasbangsa.ac.id/pkm/index.php/kesatria/article/view/164/163

[6] A. Bachir, A. Sultan, and S. S. Abu-Naser, "Predictive Modeling of Breast Cancer Diagnosis Using Neural Networks:A Kaggle Dataset Analysis," *International Journal of Academic Engineering Research*, vol. 7, no. 9, pp. 1–9, 2023, [Online]. Available: www.ijeais.org/ijaer

[7] V. Renedominick and S. Barus, "Analisis Sentimen pada Trailer Deadpool vs Wolverine Menggunakan Model Machine Learning," *Jurnal Pustaka AI (Pusat Akses Kajian Teknologi Artificial Intelligence)*, vol. 5, no. 1, pp. 01–06, 2025, doi: 10.55382/jurnalpustakaai.v5i1.892.

[8] F. Putri, Z. Awliya, and U. Budiyanto, "Analisis Sentimen Publik Twitter dengan TF-IDF dan Analysis of Public Sentiment on Twitter Using TF-IDF And," vol. 4, no. September, pp. 346–354, 2025.

[9] Z. F. Abdul Jalil, Ahmad Homaidi, "Implementasi Algoritma Support Vector Machine Untuk Klasifikasi Status Stunting Pada Balita," *G-Tech: Jurnal Teknologi Terapan*, vol. 8, no. 3, pp. 2070–2079, 2024.

[10] E. R. Lidinillah, T. Rohana, and A. R. Juwita, "Analisis sentimen twitter terhadap steam menggunakan algoritma logistic regression dan support vector machine," *TEKNOSAINS : Jurnal Sains, Teknologi dan Informatika*, vol. 10, no. 2, pp. 154–164, 2023, doi: 10.37373/tekno.v10i2.440.

[11] D. Remawati, E. Noersasongko, A. Marjuni, and Pujiono, "Mental Health Detection with TF-IDF Feature Extraction," *International Conference on Artificial Intelligence and Mechatronics System, AIMS 2024*, pp. 1–6, 2024, doi: 10.1109/AIMS61812.2024.10512480.

[12] U. Lathifah and R. Danar Dana, "Implementasi Metode Linear Regression untuk Prediksi Harga Properti Real Estate Menggunakan Rapidminer," 2024. [Online]. Available: https://www.kaggle.com/datasets/oddyvirgantara/har

[13] A. Bachir, A. Sultan, and S. S. Abu-Naser, "Predictive Modeling of Breast Cancer Diagnosis Using Neural Networks:A Kaggle Dataset Analysis," 2023. [Online]. Available: www.ijeais.org/ijaer

[14] I. D. Hartarti, I. A. Septiyani, D. A. Gultom, Y. Hendrian, and S. L. Kinanti, "Prediksi Harga Rumah di Boston Dengan Model Regresi Linear Menggunakan Python," *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 2, pp. 4250–4256, 2025, doi: 10.31004/riggs.v4i2.1210.

[15] J. A. Wibowo, V. C. Mawardi, and T. Sutrisno, "Visualisasi Word Cloud Hasil Analisis Sentimen Berbasis Fitur Layanan Aplikasi Gojek dengan Support Vector Machine," *Jurnal Serina Sains, Teknik dan Kedokteran*, vol. 2, no. 1, pp. 61–70, Mar. 2024, doi: 10.24912/jsstk.v2i1.32058.

---