**Jurnal Indonesia : Manajemen Informatika dan Komunikasi**
https://journal.stmiki.ac.id

E-ISSN: 2723-7079 | P-ISSN: 2776-8074
Vol. 6 No. 3 (2025) | September

OPEN ACCESS

https://doi.org/10.63447/jimik.v6i3.1481

# Efektivitas *Logistic Regression* dalam Analisis Sentimen Berbahasa Indonesia pada Komentar YouTube tentang Isu Ketenagakerjaan

Hamdan Santani Mulyono [1*], Usep Saprudin [2]

[1*,2] Program Studi Teknik Informatika, Universitas Dharma Wacana, Kota Metro, Provinsi Lampung, Indonesia.

*Corresponding Email*: santanihamdan@gmail.com [1*]

**Histori Artikel:**
*Dikirim* 10 Juni 2025; *Diterima dalam bentuk revisi* 2 Juli 2025; *Diterima* 10 Juli 2025; *Diterbitkan* 10 September 2025. 

## Abstrak

Studi ini mengkaji pengembangan sistem klasifikasi sentimen untuk komentar YouTube berbahasa Indonesia yang membahas permasalahan ketenagakerjaan melalui implementasi algoritma Logistic Regression. Dataset penelitian terdiri dari 2.755 komentar yang diekstrak dari video bertema "Cerita Job Seeker", dengan 1.020 komentar dilabeli secara manual ke dalam tiga kategori sentimen: positif, netral, dan negatif. Metodologi penelitian meliputi tahapan pra-pemrosesan teks, transformasi fitur menggunakan TF-IDF, pembagian data dengan stratified sampling, penanganan ketidakseimbangan kelas melalui SMOTE, dan optimasi hyperparameter menggunakan GridSearchCV. Evaluasi model menghasilkan akurasi 44% dengan distribusi performa yang bervariasi antar kelas. Kelas negatif menunjukkan performa optimal dengan F1-score 0.55, sementara kelas netral dan positif masing-masing memperoleh skor 0.34 dan 0.29. Ketidakseimbangan distribusi kelas dan karakteristik implisit komentar positif menjadi hambatan utama dalam proses klasifikasi. Hasil penelitian memperlihatkan bahwa kombinasi Logistic Regression, TF-IDF, dan SMOTE berpotensi sebagai metode baseline untuk analisis sentimen komentar media sosial berbahasa Indonesia. Meskipun demikian, pengembangan model berbasis deep learning diperlukan untuk meningkatkan akurasi dan kemampuan interpretasi nuansa linguistik. Analisis juga mengidentifikasi dominasi sentimen negatif dalam respons publik, yang mencerminkan keprihatinan masyarakat terhadap situasi ketenagakerjaan nasional.

**Kata Kunci:** Analisis Sentimen; Komentar YouTube; Logistic Regression; Ketenagakerjaan; Bahasa Indonesia.

## Abstract

This study examines the development of a sentiment classification system for Indonesian-language YouTube comments addressing employment issues through the implementation of Logistic Regression algorithm. The research dataset comprises 2,755 comments extracted from a video themed "Job Seeker Stories," with 1,020 comments manually labeled into three sentiment categories: positive, neutral, and negative. The research methodology includes text preprocessing stages, feature transformation using TF-IDF, data splitting with stratified sampling, class imbalance handling through SMOTE, and hyperparameter optimization using GridSearchCV. Model evaluation yielded 44% accuracy with varying performance distribution across classes. The negative class demonstrated optimal performance with an F1-score of 0.55, while neutral and positive classes achieved scores of 0.34 and 0.29, respectively. Class distribution imbalance and implicit characteristics of positive comments became primary obstacles in the classification process. Research findings indicate that the combination of Logistic Regression, TF-IDF, and SMOTE has potential as a baseline method for sentiment analysis of Indonesian social media comments. Nevertheless, deep learning-based model development is necessary to improve accuracy and linguistic nuance interpretation capabilities. The analysis also identified negative sentiment dominance in public responses, reflecting societal concerns regarding the national employment situation.

**Keyword:** Sentiment Analysis; YouTube Comments; Logistic Regression; Employment; Indonesian Language.

Jurnal Indonesia : Manajemen Informatika dan Komunikasi
https://journal.stmiki.ac.id
E-ISSN: 2723-7079 | P-ISSN: 2776-8074
Vol. 6 No. 3 (2025) | September
OPEN ACCESS
https://doi.org/10.63447/jimik.v6i3.1481

# 1. Pendahuluan

Tingkat pengangguran di Indonesia masih menjadi isu sosial yang signifikan. Berdasarkan data Badan Pusat Statistik, jumlah pengangguran terbuka mencapai 7,28 juta orang pada Februari 2025 (BPS, 2025), menunjukkan bahwa banyak individu usia produktif belum terserap ke dalam pasar tenaga kerja. Kondisi ini tidak hanya berdampak pada aspek ekonomi, tetapi juga memunculkan respons emosional dari masyarakat yang seringkali diekspresikan melalui platform digital. *YouTube*, sebagai salah satu media sosial berbasis video dengan tingkat interaksi tinggi, menjadi ruang bagi masyarakat untuk menyampaikan opini mereka secara terbuka. Komentar-komentar pada video bertema ketenagakerjaan, seperti video "Cerita *Job Seeker* soal Sulitnya Mendapatkan Pekerjaan" yang diunggah oleh kanal IBF tvOne, mencerminkan spektrum sentimen yang luas: mulai dari keprihatinan dan kritik, hingga harapan terhadap perbaikan kebijakan. Analisis terhadap komentar-komentar ini dapat memberikan wawasan berharga mengenai persepsi publik terhadap permasalahan ketenagakerjaan di Indonesia.

Analisis sentimen merupakan salah satu pendekatan dalam *Natural Language Processing* (NLP) yang bertujuan untuk mengidentifikasi polaritas emosi dalam teks, yakni positif, negatif, atau netral (Liu, 2022). Teknik ini banyak digunakan dalam pemantauan opini publik, evaluasi layanan, hingga pengambilan keputusan berbasis data. Namun, sebagian besar penelitian analisis sentimen di Indonesia masih terbatas pada penggunaan metode klasik seperti *Naïve Bayes* atau pendekatan berbasis leksikon, dengan sedikit eksplorasi terhadap model yang lebih fleksibel seperti *Logistic Regression*. Penelitian oleh Misrun *et al.* (2023) mengimplementasikan *Naïve Bayes* dalam klasifikasi komentar politik di *YouTube* dan memperoleh akurasi sebesar 78%. Hudha *et al.* (2022) menganalisis tayangan #MataNajwaMenantiTerawan dan menemukan bahwa mayoritas komentar bersifat netral dengan akurasi model mencapai 90,36%. Sementara itu, Sanjaya dan Lhaksmana (2020) menunjukkan efektivitas pendekatan *lexicon-based* lokal dalam memahami sentimen pada isu kabinet pemerintahan. Meskipun demikian, studi yang secara eksplisit menggunakan *Logistic Regression* untuk analisis sentimen berbahasa Indonesia, terutama dalam konteks komentar *YouTube* dan isu ketenagakerjaan, masih jarang ditemukan. *Logistic Regression* menawarkan beberapa keunggulan yang relevan untuk analisis teks pendek, seperti komentar *YouTube*. Tidak seperti *Naïve Bayes* yang mengasumsikan independensi antar fitur, *Logistic Regression* mampu menangkap hubungan linier antar kata, serta memberikan interpretasi parameter yang lebih transparan dibandingkan model berbasis *deep learning* (Birjali *et al.*, 2021). Model ini juga relatif lebih stabil dan cepat dibandingkan metode yang lebih kompleks, sehingga cocok digunakan pada data berskala menengah.

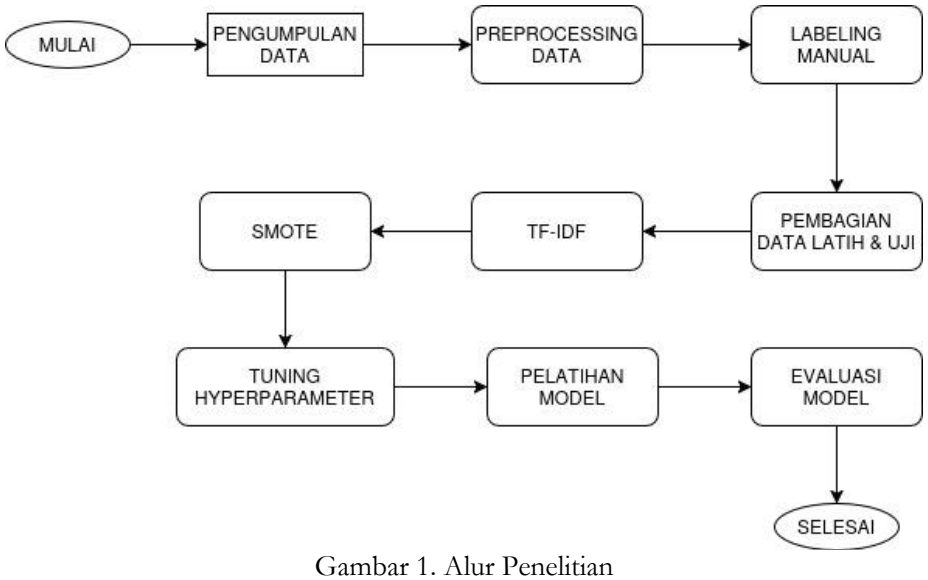
Namun, tantangan utama dalam klasifikasi komentar *YouTube* berbahasa Indonesia adalah ketidakseimbangan distribusi sentimen. Komentar negatif cenderung lebih dominan, sementara komentar positif dan netral sering kali menjadi kelas minoritas. Ketidakseimbangan ini berpotensi menurunkan performa model pada kelas minoritas, seperti yang ditunjukkan oleh studi-studi terdahulu (Gosain & Sardana, 2017). Oleh karena itu, penting untuk tidak hanya mengimplementasikan model klasifikasi, tetapi juga mengevaluasi strategi penanganan data tidak seimbang dan efektivitas model dalam konteks bahasa Indonesia.

Penelitian ini bertujuan untuk mengevaluasi efektivitas algoritma *Logistic Regression* dalam mengklasifikasikan sentimen komentar *YouTube* berbahasa Indonesia terkait isu ketenagakerjaan. Penelitian ini juga menganalisis tantangan yang timbul akibat ketidakseimbangan kelas, serta mengkaji bagaimana pendekatan ini dapat memberikan gambaran kuantitatif atas persepsi publik terhadap isu tersebut. Dengan fokus pada konteks bahasa Indonesia dan data komentar media sosial, hasil penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan metode analisis sentimen yang lebih adaptif dan aplikatif, serta menjadi referensi bagi pengambil kebijakan dan peneliti di bidang sosial digital.

Jurnal Indonesia : Manajemen Informatika dan Komunikasi
https://journal.stmiki.ac.id
E-ISSN: 2723-7079 | P-ISSN: 2776-8074
Vol. 6 No. 3 (2025) | September
OPEN ACCESS
https://doi.org/10.63447/jimik.v6i3.1481

## 2. Metode Penelitian

Penelitian ini dilakukan melalui beberapa tahapan sistematis untuk membangun model klasifikasi sentimen komentar *YouTube* berbahasa Indonesia menggunakan algoritma *Logistic Regression*. Tahapan-tahapan tersebut meliputi pengumpulan data, pelabelan sentimen, prapemrosesan teks, transformasi fitur dengan TF-IDF, penanganan data tidak seimbang, pelatihan model, dan evaluasi performa. Alur metodologi penelitian ini digambarkan pada Gambar 1.




Gambar 1. Alur Penelitian

### 2.1 Pengumpulan Data

Data dikumpulkan dari platform *YouTube* dengan menggunakan *YouTube Data API v3*, melalui *library* googleapiclient.discovery dari *Python*. Video yang menjadi objek penelitian berjudul "Cerita *Job Seeker* soal Sulitnya Mendapatkan Pekerjaan", diunggah oleh kanal IBF tvOne, dengan ID video HLOErnZKipI. Komentar utama dan balasan diambil menggunakan sistem *pagination*, dengan batas maksimum 100 komentar atau balasan per halaman. Proses *crawling* berhasil mengumpulkan sebanyak 2.755 komentar, yang mencakup komentar utama dan balasan. Data tersebut kemudian disimpan dalam format CSV untuk keperluan pemrosesan lebih lanjut.

### 2.2 Prapemrosesan Data

Data komentar yang telah dikumpulkan selanjutnya melalui proses prapemrosesan (*text preprocessing*) (Rianto *et al.*, 2021) untuk mempersiapkan teks agar dapat digunakan oleh model klasifikasi. Tahapan ini terdiri dari:

1) *Cleaning*
   Menghapus karakter non-alfabet, angka, tanda baca, simbol, dan *emoji*.
2) *Case Folding*
   Mengubah seluruh huruf menjadi huruf kecil.
3) *Tokenizing*
   Memecah komentar menjadi *token* (kata).
4) *Stopword Removal*
   Menghapus kata-kata umum yang tidak memiliki makna signifikan.
5) *Stemming*
   Mengembalikan kata ke bentuk dasarnya menggunakan *stemmer* Bahasa Indonesia (misalnya Sastrawi).

Jurnal Indonesia : Manajemen Informatika dan Komunikasi
https://journal.stmiki.ac.id
E-ISSN: 2723-7079 | P-ISSN: 2776-8074
Vol. 6 No. 3 (2025) | September
OPEN ACCESS
https://doi.org/10.63447/jimik.v6i3.1481

Hasil prapemrosesan disimpan dalam kolom Cleaned_Comment.

### 2.3 Pelabelan Manual

Dari keseluruhan data, sebanyak 1.020 komentar dilabeli secara manual oleh penulis sebagai satu-satunya *annotator*. Label yang digunakan terdiri dari tiga kelas: positif, netral, dan negatif, berdasarkan polaritas emosi yang teridentifikasi dalam setiap komentar. Proses pelabelan ini bersifat subjektif, sehingga tidak dilakukan pengukuran *inter-annotator agreement* (Liu, 2022). Meskipun demikian, pelabelan dilakukan dengan mempertimbangkan konteks kalimat dan makna implisit dalam komentar. *Dataset* hasil pelabelan disimpan dalam file manual_labeling_preprocessed_with_sentiment.csv dan hanya data yang memiliki label yang digunakan untuk pelatihan model.

### 2.4 Transformasi Teks dengan TF-IDF

Teks yang telah dibersihkan diubah menjadi vektor numerik menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*) untuk digunakan sebagai input model (Ramos, 2003). Parameter yang digunakan dalam TfidfVectorizer adalah:

1) ngram_range = (1, 2) untuk mempertimbangkan *unigram* dan *bigram*,
2) max_df = 0.85 untuk menghindari fitur yang terlalu sering muncul.

TF-IDF dipilih karena mampu menangkap pentingnya suatu kata atau frasa dalam konteks dokumen tertentu, dan terbukti efektif untuk teks pendek seperti komentar media sosial.

### 2.5 Pembagian Data Latih dan Uji

Data yang telah diproses dibagi menjadi data latih (80%) dan data uji (20%) menggunakan fungsi train_test_split dari *Scikit-learn*, dengan metode *stratified sampling* untuk memastikan distribusi proporsi kelas tetap terjaga. Setelah penyaringan label yang kosong, jumlah data yang digunakan dalam pelatihan model adalah 1.020 komentar, dengan 816 data latih dan 204 data uji.

### 2.6 Penanganan Ketidakseimbangan Data

Distribusi awal data latih menunjukkan ketidakseimbangan kelas yang signifikan, dengan 386 komentar negatif, 280 netral, dan 150 positif. Ketidakseimbangan ini dapat menyebabkan model bias terhadap kelas mayoritas dan mengabaikan kelas minoritas. Oleh karena itu, diterapkan teknik SMOTE (*Synthetic Minority Oversampling Technique*) (Elreedy & Atiya, 2019) pada data latih untuk menyintesis data baru pada kelas minoritas. Hasil penyeimbangan menghasilkan jumlah data yang merata, yaitu 386 komentar untuk masing-masing kelas (negatif, netral, dan positif). Proses SMOTE diterapkan hanya pada data latih, untuk menghindari *data leakage* ke dalam data uji.

### 2.7 *Tuning Hyperparameter*, Pelatihan, dan Evaluasi Model

Setelah penyeimbangan data, dilakukan proses *tuning hyperparameter* terhadap algoritma *Logistic Regression* menggunakan GridSearchCV (Passos & Mishra, 2022) dengan 5-*fold cross-validation*. Parameter yang diuji meliputi: C: [0.1, 1, 10], penalty: ['l1', 'l2'], solver: ['liblinear']. Kombinasi terbaik diperoleh pada: C = 10, penalty = 'l2', solver = 'liblinear'. Model terbaik ini digunakan dalam *pipeline* klasifikasi dan diuji menggunakan 204 data uji. Evaluasi dilakukan menggunakan metrik:

1) *Accuracy*
2) *Precision*
3) *Recall*
4) *F1-Score*
5) *Confusion Matrix*

Hasil evaluasi menunjukkan bahwa model memiliki akurasi 44%, dengan *F1-score* tertinggi pada kelas negatif (0,55), dan *F1-score* yang meningkat pada kelas positif (0,29) dan netral (0,34). Ini menunjukkan

Jurnal Indonesia : Manajemen Informatika dan Komunikasi
https://journal.stmiki.ac.id
E-ISSN: 2723-7079 | P-ISSN: 2776-8074
Vol. 6 No. 3 (2025) | September
OPEN ACCESS
https://doi.org/10.63447/jimik.v6i3.1481

bahwa penerapan SMOTE dan *tuning* berhasil meningkatkan representasi dan keseimbangan performa model, meskipun tantangan dalam mengklasifikasikan komentar positif masih terlihat.

## 3. Hasil dan Pembahasan

### 3.1 Hasil

Hasil implementasi algoritma *Logistic Regression* dalam klasifikasi sentimen komentar *YouTube* berbahasa Indonesia yang berkaitan dengan isu ketenagakerjaan. Penilaian dilakukan melalui dua tahap utama, yakni sebelum dan setelah penyeimbangan data menggunakan SMOTE (Xu *et al*., 2020) serta penerapan *tuning hyperparameter*. Evaluasi didasarkan pada metrik presisi, *recall*, *F1-score*, dan *confusion matrix*.

#### 3.1.1 Distribusi Label dan Tantangan *Imbalanced Data*

Dari total 2.755 komentar yang dikumpulkan, sebanyak 1.020 komentar berhasil dilabeli secara manual oleh penulis. Distribusi sentimen menunjukkan ketimpangan yang cukup besar:

1) Negatif: 482 komentar
2) Netral: 350 komentar
3) Positif: 188 komentar

Setelah dilakukan pembagian data menggunakan *stratified sampling* (80:20), data latih terdiri dari 816 data, sedangkan data uji sebanyak 204 data. Distribusi data latih tetap menunjukkan ketidakseimbangan: 386 negatif, 280 netral, dan 150 positif. Hal ini berpotensi menyebabkan model bias terhadap kelas mayoritas (Dablain *et al*., 2022).

#### 3.1.2 Prapemrosesan Data

Dalam langkah prapemrosesan data mencakup beberapa langkah seperti *cleaning*, *case folding*, pelabelan, *tokenizing*, *stopword removal*, dan *stemming*. Hasil dari prapemrosesan data dalam Tabel 1 berikut.

Tabel 1. Contoh Hasil Prapemrosesan Data

| Prapemrosesan | Input | Output |
|---|---|---|
| *Cleaning* | Sama aja..kenyataan yg lulus sarja jadi sopir | Sama aja kenyataan yg lulus sarja jadi sopir |
| *Case folding* | Sama aja kenyataan yg lulus sarja jadi sopir | sama aja kenyataan yg lulus sarja jadi sopir |
| *Tokenizing* | sama aja kenyataan yg lulus sarja jadi sopir | ['sama', 'aja', 'kenyataan', 'yg', 'lulus', 'sarja', 'jadi', 'sopir'] |
| *Stopword removal* | ['masalah', 'di', 'indonesia', 'gak', 'pernah', 'selesai', 'karena', 'dr', 'riset', 'yg', 'saya', 'lakukan', 'setiap', 'maslah', 'hanya', 'di', 'obrilin', 'sj', 'tidak', 'berusaha', 'membuat', 'solusi'] | ['masalah', 'indonesia', 'gak', 'pernah', 'selesai', 'dr', 'riset', 'yg', 'lakukan', 'maslah', 'obrilin', 'sj', 'berusaha', 'membuat', 'solusi'] |
| *Stemming* | ['masalah', 'indonesia', 'gak', 'pernah', 'selesai', 'dr', 'riset', 'yg', 'lakukan', 'maslah', 'obrilin', 'sj', 'berusaha', 'membuat', 'solusi'] | ['masalah', 'indonesia', 'gak', 'pernah', 'selesai', 'dr', 'riset', 'yg', 'laku', 'mas', 'obrilin', 'sj', 'usaha', 'buat', 'solusi'] |

Kemudian pada proses pelabelan menghasilkan distribusi sentimen sebagai berikut:

Jurnal Indonesia : Manajemen Informatika dan Komunikasi
https://journal.stmiki.ac.id
E-ISSN: 2723-7079 | P-ISSN: 2776-8074
Vol. 6 No. 3 (2025) | September
OPEN ACCESS
https://doi.org/10.63447/jimik.v6i3.1481

Tabel 2. Hasil Distribusi Sentimen

| Sentimen | Jumlah Komentar |
|---|---|
| Negatif | 482 |
| Netral | 350 |
| Positif | 188 |

Distribusi ini menunjukkan bahwa mayoritas komentar bersifat negatif, mencerminkan kekhawatiran dan ketidakpuasan masyarakat terhadap isu ketenagakerjaan yang dibahas dalam video.

3.1.3 Klasifikasi dengan *Logistic Regression*

Data yang telah diproses dibagi menjadi data latih (80%) dan data uji (20%) dengan *stratified sampling*. Model klasifikasi dibangun dalam bentuk *pipeline* yang terdiri dari: • Transformasi teks menggunakan TF-IDF *Vectorizer* (ngram_range=(1,2), max_df=0.85) • Algoritma *Logistic Regression* sebagai *classifier* utama

3.1.4 *Tuning* Model (*GridSearchCV*)

Agar diperoleh hasil klasifikasi yang optimal, dilakukan *tuning hyperparameter* menggunakan *GridSearchCV* (Passos & Mishra, 2022) dengan 5-*fold cross-validation*. Parameter yang diujikan meliputi nilai regulasi (C), jenis penalti (l1 atau l2), dan jenis *solver*. Hasil kombinasi terbaik:

1) max_df = 0.85
2) ngram_range = (1, 2)
3) penalty = 'l2'
4) C = 10
5) solver = 'liblinear'

Kombinasi ini menghasilkan model terbaik yang digunakan dalam proses evaluasi.

3.1.5 Evaluasi Model

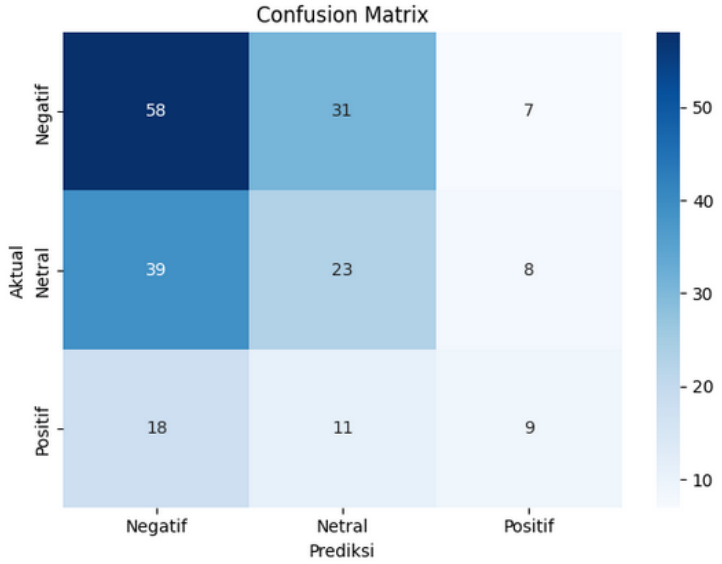
Model yang telah di-*tuning* diuji menggunakan 204 data uji. Hasil evaluasi disajikan dalam Tabel 3.

Tabel 3. Hasil Evaluasi Model *Logistic Regression*

| Sentimen | *Precision* | *Recall* | *F1-Score* | *Support* |
|---|---|---|---|---|
| Negatif | 0,50 | 0,60 | 0,55 | 96 |
| Netral | 0,35 | 0,33 | 0,34 | 70 |
| Positif | 0,38 | 0,24 | 0,29 | 38 |
| Akurasi | - | - | 0,44 | 204 |

Berdasarkan hasil evaluasi pada Tabel 3, terlihat bahwa model *Logistic Regression* menunjukkan performa yang bervariasi untuk setiap kelas sentimen. Kelas negatif memperoleh nilai *F1-score* tertinggi (0,55) dengan *recall* yang cukup baik (0,60), menunjukkan bahwa model relatif berhasil mengidentifikasi komentar bersentimen negatif. Hal ini dapat dipahami mengingat kelas negatif memiliki jumlah data terbanyak dalam *dataset* pelatihan. Sebaliknya, kelas positif menunjukkan performa terendah dengan *F1-score* 0,29 dan *recall* hanya 0,24, mengindikasikan kesulitan model dalam mengenali sentimen positif. Kelas netral berada di posisi tengah dengan *F1-score* 0,34, yang mencerminkan tantangan dalam mengklasifikasikan komentar yang tidak memiliki polaritas sentimen yang jelas. Visualisasi hasil prediksi ditampilkan melalui *confusion matrix* pada Gambar 2.

Jurnal Indonesia : Manajemen Informatika dan Komunikasi
https://journal.stmiki.ac.id
E-ISSN: 2723-7079 | P-ISSN: 2776-8074
Vol. 6 No. 3 (2025) | September
OPEN ACCESS
https://doi.org/10.63447/jimik.v6i3.1481




Gambar 2. *Confusion Matrix*

### 3.2 Pembahasan

Model menunjukkan performa terbaik pada kelas negatif dengan *F1-score* sebesar 0,55, yang dapat dijelaskan oleh dominasi jumlah komentar negatif dalam data pelatihan. Sebaliknya, performa pada kelas positif masih rendah (*F1-score*: 0,29; *recall*: 0,24), yang menunjukkan banyak komentar positif tidak dikenali dengan baik oleh model. Hal ini mencerminkan tantangan dalam menangani ketidakseimbangan kelas dan ekspresi positif yang sering kali bersifat implisit dalam komentar berbahasa Indonesia. Kelas netral juga menunjukkan hasil sedang dengan *F1-score* 0,34, mengindikasikan adanya ambiguitas semantik yang umum ditemukan pada teks informal di media sosial. Rendahnya performa pada kelas netral dapat disebabkan oleh sifat komentar yang seringkali mengandung campuran sentimen atau menggunakan bahasa yang tidak eksplisit dalam menyatakan polaritas emosi. Dengan akurasi keseluruhan sebesar 44%, model *Logistic Regression* yang dikombinasikan dengan TF-IDF dan SMOTE terbukti mampu melakukan klasifikasi dasar terhadap sentimen komentar *YouTube*. Namun, keterbatasannya dalam menangkap sentimen minoritas menunjukkan perlunya pendekatan lanjutan. Ke depan, pengembangan model berbasis *deep learning* seperti LSTM atau BERT direkomendasikan untuk meningkatkan pemahaman konteks dan nuansa bahasa, serta menghasilkan prediksi yang lebih akurat dan seimbang (Devlin *et al.*, 2019). Temuan ini juga mencerminkan dominasi opini negatif masyarakat terhadap isu ketenagakerjaan, yang dapat menjadi indikator sosial bagi pembuat kebijakan dan analis opini publik. Tingginya proporsi sentimen negatif dalam komentar menunjukkan adanya ketidakpuasan masyarakat terhadap kondisi ketenagakerjaan di Indonesia, yang perlu menjadi perhatian serius bagi pemerintah dalam merumuskan kebijakan yang lebih responsif terhadap kebutuhan pencari kerja.

## 4. Kesimpulan

Penelitian ini berhasil membangun sistem klasifikasi sentimen pada komentar *YouTube* terkait isu ketenagakerjaan menggunakan algoritma *Logistic Regression*. Proses dilakukan melalui tahapan pengumpulan data, pelabelan manual terhadap 1.020 komentar, prapemrosesan teks, transformasi fitur dengan TF-IDF, penyeimbangan data menggunakan SMOTE, *tuning hyperparameter* dengan *GridSearchCV*, serta evaluasi model. Hasil evaluasi menunjukkan bahwa model memiliki akurasi sebesar 44%, dengan performa terbaik pada kelas negatif (*F1-score*: 0,55), sedangkan performa pada

E-ISSN: 2723-7079 | P-ISSN: 2776-8074
Vol. 6 No. 3 (2025) | September
OPEN ACCESS
https://doi.org/10.63447/jimik.v6i3.1481

kelas netral (*F1-score*: 0,34) dan positif (*F1-score*: 0,29) masih tergolong rendah. Hal ini mencerminkan pengaruh signifikan dari ketidakseimbangan distribusi label, serta tantangan dalam mengidentifikasi komentar positif yang cenderung implisit dan kontekstual. Penerapan *GridSearchCV* membantu dalam memperoleh konfigurasi optimal (C=10, penalty='l2', solver='liblinear'), namun keterbatasan dalam memahami nuansa bahasa menunjukkan bahwa pendekatan ini masih bersifat dasar. Secara keseluruhan, *Logistic Regression* yang dikombinasikan dengan TF-IDF dan SMOTE dapat digunakan sebagai *baseline* untuk analisis sentimen berbahasa Indonesia, namun pengembangan lebih lanjut dengan model yang lebih kompleks seperti LSTM atau BERT sangat disarankan untuk meningkatkan akurasi dan generalisasi. Temuan ini juga memperlihatkan bahwa opini publik terkait ketenagakerjaan di Indonesia lebih banyak bersentimen negatif, yang dapat menjadi bahan pertimbangan bagi pemangku kebijakan, lembaga ketenagakerjaan, serta peneliti dalam memahami respons masyarakat terhadap isu sosial dan ekonomi secara digital.

## 5. Daftar Pustaka

Ash, S., & Surya, A. (2022). Analisis sentimen masyarakat terhadap kebijakan vaksinasi COVID-19 pada media sosial Twitter menggunakan metode logistic regression. *Analisis Sentimen Masyarakat Terhadap Kebijakan Vaksinasi Covid-19 Pada Media Sosial Twitter Menggunakan Metode Logistic Regression*, *3*(2), 99–106. https://doi.org/10.37859/coscitech.v3i2.3836

Badan Pusat Statistik. (2025, Februari). BPS: Jumlah pengangguran naik jadi 7,28 juta orang per Februari 2025. *Tempo*. https://www.tempo.co/ekonomi/bps-jumlah-pengangguran-naik-jadi-7-28-juta-orang-per-februari-2025-1344338

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134. https://doi.org/10.1016/j.knosys.2021.107134

Dablain, D., Krawczyk, B., & Chawla, N. V. (2022). DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(9), 6390–6404. https://doi.org/10.1109/TNNLS.2021.3136503

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information Sciences*, *505*, 32–64. https://doi.org/10.1016/j.ins.2019.07.070

Gosain, A., & Sardana, S. (2017). Handling class imbalance problem using oversampling techniques: A review. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 79–85). IEEE. https://doi.org/10.1109/ICACCI.2017.8125820

Hudha, M., Supriyati, E., & Listyorini, T. (2022). Analisis sentimen pengguna YouTube terhadap tayangan #matanajwamenantiterawan dengan metode naïve bayes classifier. *JIKO (Jurnal Informatika dan Komputer)*, *5*(1), 1–6.

Liu, B. (2022). *Sentiment analysis and opinion mining*. Springer Nature.

E-ISSN: 2723-7079 | P-ISSN: 2776-8074
Vol. 6 No. 3 (2025) | September
OPEN ACCESS
https://doi.org/10.63447/jimik.v6i3.1481

Misrun, C. A., Haerani, E., Fikry, M., & Budianita, E. (2023). Analisis sentimen komentar YouTube terhadap Anies Baswedan sebagai bakal calon presiden 2024 menggunakan metode naive bayes classifier. *Jurnal Coscitech (Computer Science and Information Technology)*, *4*(1), 207–215. https://doi.org/10.37859/coscitech.v4i1.4790

Passos, D., & Mishra, P. (2022). A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks. *Chemometrics and Intelligent Laboratory Systems*, *223*, 104520. https://doi.org/10.1016/j.chemolab.2022.104520

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning* (pp. 29–48). Citeseer.

Rianto, Mutiara, A. B., Wibowo, E. P., & Santosa, P. I. (2021). Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. *Journal of Big Data*, *8*, 1–16. https://doi.org/10.1186/s40537-021-00413-1

Sanjaya, G., & Lhaksmana, K. M. (2020). Analisis sentimen komentar YouTube tentang terpilihnya menteri kabinet Indonesia maju menggunakan lexicon based. *eProceedings of Engineering*, *7*(3).

Xu, Z., Shen, D., Nie, T., & Kou, Y. (2020). A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. *Journal of Biomedical Informatics*, *107*, 103465. https://doi.org/10.1016/j.jbi.2020.103465