

Sentiment Analysis of YouTube Social Media on Lumpur Lapindo Tourism Using the K-Nearest Neighbor (K-NN) Algorithm and Term Frequency – Inverse Document Frequency

[Analisis Sentimen pada Media Sosial Youtube Terhadap Pariwisata Lumpur Lapindo dengan Menggunakan Algoritma K-Nearest Neighbor (K-NN) dan Term Frequency – Invers Document Frequency]

Muhammad Iqbal Nahariqi¹⁾, Yulian Findawati ^{*,2)}

¹⁾ Program Studi Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia

²⁾ Program Studi Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia

*Email Penulis Korespondensi: yulianfindawati@umsida.ac.id

Abstract. Digital transformation in the current industrial era has had a significant impact on various aspects of human life, including communication patterns and social interactions. One rapidly growing digital platform is YouTube, which serves as a medium for sharing information in asynchronous video formats. This platform provides a space for users to interact through the comment feature, allowing the public to express opinions, criticisms, and suggestions regarding the content presented. To assess public perception of tourism in the Lapindo Mud area, sentiment analysis was conducted on YouTube user comments. This study utilizes the K-Nearest Neighbor (K-NN) algorithm for opinion classification. The research stages include data collection, text preprocessing to obtain clean and structured data, and word weighting using the Term Frequency–Inverse Document Frequency (TF-IDF) method. All these processes aim to produce accurate sentiment classification. The data used in this study is primary data obtained directly from the YouTube platform.

Keywords - Youtube, , Classification, Sentiment analysis, K-Nearest Neighbor(K-NN)

Abstrak. Transformasi digital pada era industri sekarang telah membawa dampak yang signifikan terhadap berbagai aspek kehidupan manusia, termasuk dalam pola komunikasi dan interaksi sosial. Salah satu platform digital yang berkembang pesat adalah YouTube, yang berfungsi sebagai media berbagi informasi dalam format video secara asinkron. Platform ini menyediakan ruang bagi pengguna untuk berinteraksi melalui fitur komentar, sehingga memungkinkan masyarakat menyampaikan pendapat, kritik, maupun saran terhadap konten yang ditayangkan. Guna menilai persepsi publik terhadap eksistensi pariwisata di kawasan Lumpur Lapindo, dilakukan analisis sentimen terhadap komentar pengguna YouTube. Penelitian ini memanfaatkan algoritma K-Nearest Neighbor (K-NN) untuk proses klasifikasi opini. Tahapan penelitian mencakup pengumpulan data, pra-pemrosesan teks untuk memperoleh data yang bersih dan terstruktur, serta pembobotan kata menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). Seluruh proses tersebut diarahkan untuk menghasilkan klasifikasi sentimen yang akurat. Adapun data yang digunakan merupakan data primer yang diperoleh langsung dari platform YouTube.

Kata Kunci - Youtube, Klasifikasi, Analisis sentimen, K-Nearest Neighbor (K-NN), Term Frequency–Invers Document Frequency (TF-IDF)

I. PENDAHULUAN

Kemajuan teknologi pada era industri sekarang telah menyebabkan transformasi besar dalam berbagai dimensi aktivitas manusia, khususnya dalam kehidupan masyarakat secara umum. Media sosial YouTube merupakan salah satu platform berbasis web dan aplikasi yang sangat diminati, dengan internet sebagai sarana utama dalam penyajian konten videonya. Platform ini memberikan keleluasaan bagi pengguna untuk mengunggah video, memberikan komentar, serta berinteraksi dengan berbagai konten. Walaupun telah berkembang menjadi salah satu repositori video terbesar di dunia, tingkat popularitasnya juga membuka peluang terjadinya penyalahgunaan [1].

Untuk mendukung pengembangan aplikasi yang terintegrasi dengan sistem YouTube, tersedia layanan **YouTube API**. Layanan ini berfungsi sebagai jembatan bagi pengembang untuk mengakses dan mengelola konten video secara langsung dari platform tersebut. Dengan memanfaatkan API ini, pengembang memperoleh fleksibilitas dalam

menciptakan aplikasi dan layanan yang lebih kompleks, sekaligus meningkatkan kualitas interaksi dan pengalaman pengguna secara keseluruhan [1].

Untuk mengetahui adanya dampak lumpur lapindo maka dilakukan Analisis sentimen terhadap pariwisata lumpur lapindo. Analisis sentimen merupakan proses pengolahan opini seseorang yang mencerminkan persepsi terhadap suatu merek, situasi, atau topik tertentu. Pendekatan ini umumnya mengelompokkan ekspresi tersebut ke dalam tiga kategori Utama, yaitu positif dan negatif yang menunjukkan tingkat kepuasan atau ketidakpuasan terhadap suatu hal [2].

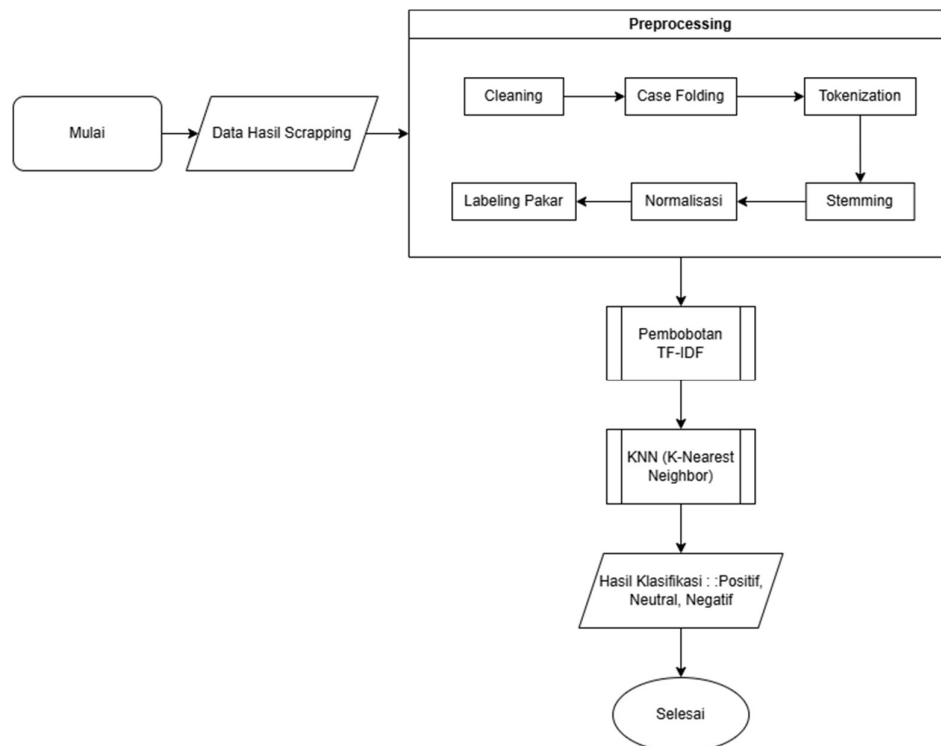
Proses klasifikasi bertujuan untuk membangun Model atau fungsi ini berperan dalam mengenali serta membedakan kelompok-kelompok data yang ada. Fungsinya adalah untuk memprediksi kategori dari data baru yang sebelumnya belum diklasifikasikan. Dalam prosesnya, klasifikasi menghasilkan model yang mampu mengelompokkan data berdasarkan pola atau aturan tertentu. Model tersebut dapat berupa logika kondisional seperti 'jika-maka', struktur pohon keputusan, ataupun representasi matematis lainnya[3].

Penelitian Terdahulu

Studi yang dilakukan oleh Muhammad Hafizh Mahendra, Danang Triantoro Murdiansyah, dan Kemas Muslim Lhaksana (2023) berjudul *"Analisis Sentimen Tweet COVID-19 Menggunakan Metode K-Nearest Neighbors dengan Ekstraksi Fitur TF-IDF dan CountVectorizer"* ini membuat sistem klasifikasi dalam analisis sentimen terhadap tweet terkait COVID-19 dilakukan dengan menerapkan metode K-Nearest Neighbor (KNN), serta menggunakan TF-IDF dan CountVectorizer sebagai teknik ekstraksi fitur. Hasil pengujian menunjukkan bahwa TF-IDF dengan 10.000 fitur dan rasio data 70:30 menghasilkan akurasi tertinggi sebesar 74,4%. Rasio data yang lebih besar untuk pelatihan tidak selalu membuat model bekerja lebih baik. Jika dibandingkan dengan CountVectorizer, metode TF-IDF menunjukkan performa akurasi yang lebih unggul, sehingga lebih cocok digunakan dalam penelitian ini. Sistem juga dapat memprediksi data dengan benar pada kategori positif sebesar 62,5% dan kategori negatif sebesar 85,5%. [4].

II. METODOLOGI

Tahapan-tahapan dalam metode penelitian ini dijelaskan secara rinci melalui flowchart berikut :



Gambar 2.1. Flowchart Metode Penelitian

Tahap awal dari sistem yaitu diawali dengan *Scrapping* komentar/ulasan dari sosial media Youtube, setelah itu dilanjutkan dengan Preprocessing, pada proses ini akan didapatkan *term* dari sebuah data, selanjutnya akan

Copyright © Universitas Muhammadiyah Sidoarjo. This preprint is protected by copyright held by Universitas Muhammadiyah Sidoarjo and is distributed under the Creative Commons Attribution License (CC BY). Users may share, distribute, or reproduce the work as long as the original author(s) and copyright holder are credited, and the preprint server is cited per academic standards.

Authors retain the right to publish their work in academic journals where copyright remains with them. Any use, distribution, or reproduction that does not comply with these terms is not permitted.

dilakukan pembobotan dengan menggunakan metode N-gram, dari proses pembobotan ini didapatkan matriks dari sebuah data yang akan dilakukan perhitungan pada proses selanjutnya yaitu menggunakan metode KNN (K-Nearest Neighbor). Untuk menentukan klasifikasi dilakukan perhitungan dengan menggunakan metode Cosine Similarity dan didapatkan hasilnya.

Scrapping

Web scrapping merupakan proses ekstraksi data dari dokumen semi-terstruktur yang terdapat pada halaman web, dengan tujuan mengambil informasi tertentu secara selektif. Teknik ini memungkinkan pengumpulan data dari internet yang selanjutnya dapat disimpan dalam format file maupun basis data untuk keperluan analisis lebih lanjut. Web scrapping dikenal sebagai metode yang efisien dan andal dalam pengumpulan data berskala besar (big data), serta memiliki kemampuan menjangkau informasi yang tidak selalu dapat ditemukan melalui mesin pencari konvensional seperti Google Search. Untuk tahapan selanjutnya dilakukan dengan *Preprocessing* [5].

Preprocessing

Text processing merupakan bagian dari teknik text mining yang bertujuan untuk mengubah data teks yang tidak terstruktur menjadi format yang lebih terorganisir, sehingga dapat disimpan dan dikelola dalam basis data [6].

1. Cleaning

Langkah ini bertujuan untuk menghapus atribut yang kurang relevan terhadap analisis, guna mendukung efisiensi serta meningkatkan akurasi pengolahan data. Sebagaimana namanya, tahap ini berfokus pada pembersihan elemen-elemen yang dianggap mengganggu atau tidak penting, guna mempersiapkan data untuk proses analisis lanjutan. Selain itu, dilakukan juga normalisasi terhadap kata-kata yang mengalami kesalahan penulisan. Elemen-elemen yang umumnya dihapus meliputi URL, tagar (#), dan mention (@) [7].

2. Case Folding

Pada tahap ini, seluruh teks diubah menjadi bentuk yang seragam, biasanya dalam huruf kecil (lowercase). Proses ini bertujuan untuk menyamakan representasi kata yang memiliki perbedaan dalam penggunaan huruf kapital, guna menjaga konsistensi dalam analisis data teks, seperti “Budi” dan “budi”, dapat dikenali sebagai entitas yang sama dalam analisis kemiripan dokumen. Sebagai contoh, kalimat “Budi sedang bermain komputer” akan diubah menjadi “budi sedang bermain komputer”. Proses ini dikenal sebagai *case folding* dan penting dilakukan agar perhitungan kemiripan antar kata atau dokumen menjadi lebih akurat [8].

3. Tokenization

Tokenisasi merupakan proses awal dalam pengolahan teks, di mana dokumen yang terdiri dari beberapa kalimat dipisahkan menjadi elemen-elemen kata yang dikenal sebagai token. Selain itu Tokenisasi mencakup pemisahan urutan string menjadi frasa, kata kunci, kata, simbol, yang disebut token [8].

4. Stemming

Langkah ini dilakukan untuk mengkonversi setiap kata ke bentuk dasarnya dengan cara menghilangkan imbuhan, sehingga mempermudah proses analisis teks, baik yang berada di awal (prefiks) maupun di akhir (sufiks) kata [11].

5. Normalisasi

Normalisasi kata merupakan proses pengubahan kata tidak baku atau tidak standar menjadi bentuk yang sesuai dengan kaidah tata bahasa Indonesia yang merujuk pada standar yang ditetapkan dalam Kamus Besar Bahasa Indonesia (KBBI). Proses ini berfungsi untuk menyampaikan gagasan secara lebih jelas dengan menyesuaikan format teks agar sesuai dengan tujuan analisis tertentu. Normalisasi teks menjadi penting karena dalam sebuah dokumen atau bacaan sering ditemukan kata-kata yang sulit dipahami, baik karena perbedaan bahasa, penggunaan istilah yang tidak umum, maupun makna yang menyimpang dari konteks aslinya [9].

TF - IDF

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan salah satu metode pembobotan kata yang mengkombinasikan dua elemen penting, yakni seberapa sering suatu istilah muncul dalam sebuah dokumen (term frequency), serta seberapa langka istilah tersebut dalam keseluruhan kumpulan dokumen (inverse document frequency). Semakin sering suatu kata muncul dalam dokumen tertentu, semakin besar kontribusinya terhadap pemaknaan dokumen tersebut. Sebaliknya, semakin jarang istilah tersebut ditemukan di seluruh dokumen, semakin tinggi bobotnya dalam proses analisis, karena dianggap lebih spesifik dan informatif [10].

K- Nearest Neighbor (K-NN)

Algoritma K-Nearest Neighbor (K-NN) termasuk dalam kelompok metode klasifikasi berbasis kedekatan, di mana penentuan kelas suatu data didasarkan pada jarak terdekat dengan sejumlah tetangga terdekat dalam ruang fitur pembelajaran mesin (*machine learning*) yang bersifat *non-parametrik* dan termasuk dalam kategori *lazy learner*. Sifat *non-parametrik* mengindikasikan bahwa algoritma ini tidak mengasumsikan bentuk distribusi tertentu pada data yang digunakan. Dengan demikian, tidak terdapat parameter tetap yang perlu diestimasi dalam model, terlepas dari ukuran data yang dianalisis. Sebagai metode *lazy learning*, KNN tidak membangun model selama proses pelatihan, melainkan

menyimpan seluruh data latih dan melakukan klasifikasi atau prediksi hanya saat proses pengujian. Akibatnya, tahap pelatihan berlangsung lebih cepat, sementara tahap pengujian cenderung memerlukan waktu dan sumber daya komputasi yang lebih besar. Prinsip dasar dari algoritma ini adalah bahwa objek-objek yang memiliki karakteristik serupa akan berada dalam jarak yang berdekatan dalam ruang fitur, sehingga proses klasifikasi dilakukan dengan mengidentifikasi sejumlah k tetangga terdekat dari data uji [11].

III. HASIL DAN PEMBAHASAN

A. Analisis data

Dataset yang digunakan dalam penelitian ini terdiri dari sekumpulan data yang telah diproses dan diklasifikasikan sesuai dengan tujuan analisis berjumlah 520 komentar/ulasan digunakan dalam penelitian ini terdiri dari 6 atribut, yaitu *NAMA*, *KOMENTAR*, *L_Pakar1*, *L_Pakar2*, *L_Pakar3*, dan *L_FIX*. Namun, hanya lima fitur yang dimanfaatkan dalam implementasi program, yaitu *KOMENTAR*, *L_Pakar1*, *L_Pakar2*, *L_Pakar3*, dan *L_FIX*. Adapun tiga pakar yang memberikan label merupakan mahasiswa dari Program Studi Psikologi di Universitas XXX, masing-masing diidentifikasi sebagai Pakar 1, Pakar 2, dan Pakar 3. Dataset yang digunakan dijelaskan secara rinci pada Tabel 1, bersama dengan deskripsi dari setiap fitur yang terdapat di dalamnya.

- Nama : Nama individu atau kelompok yang memberikan komentar terhadap unggahan video youtube “Pariwisata Lumpur Lapindo”
- Komentar : Kolom yang berisi tanggapan, kritik, pernyataan, ataupun saran terhadap pariwisata lumpur lapindo
- L_PAKAR 1 : Hasil pelabelan komentar yang dilakukan oleh Pakar 1
- L_PAKAR 2 : Hasil pelabelan komentar yang dilakukan oleh Pakar 2
- L_PAKAR 3 : Hasil pelabelan komentar yang dilakukan oleh Pakar 3
- L_FIX : Pelabelan yang diambil berdasarkan hasil voting yang dilakukan oleh ketiga pakar.
- Kelas dataset tersebut terdiri dari dua kelas, yakni Positif dan Negatif. Kelas Positif berjumlah 169 data, sedangkan kelas Negatif terdiri dari 351 data..

TABEL 3.1 Contoh dataset yang digunakan

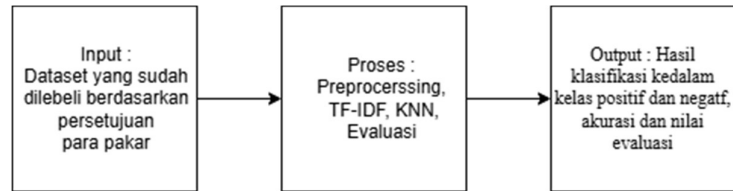
Nama	Komentar	L_Pakar1	L_Pakar2	L_Pakar3	L_FIX
Niko_Channel	Yang kemaren banyak tanya tentang sejarah dan kisah Lumpur Lapindo, ini sudah saya buat video khusus cerita asli dari pak Aksan loh, ini link videonya: https://youtu.be/6YVVGIxCfU Ada juga yang tanya di mana pembuangan lumpur lapindo, ini link video nya: https://youtu.be/bzHwwkXamuQ Terima kasih kepada pak Aksan yang telah membantu saya dalam membuat konten ini..	Negatif	Negatif	Negatif	Negatif
Arie J	Wah Pak Aksan ini bukan tukang ojek biasa.. benar2 seperti pemandu wisata. Penjelasan detail dan informatif. Beliau ini orang cerdas dari bahasanya yang tertata rapi dan caranya menjawab pertanyaan2 mas niko dengan tepat sasaran, lugas, jelas. Sayang, beliau cuma mematok harga 50rb untuk satu kali perjalanan	Positif	Positif	Negatif	Positif

	berangkat-pulang. Padahal harga segitu gak sebanding dengan semua informasi yg diberikan :((Semoga lancar terus ya rejekinya Pak.. mudah2an ada kesempatan bagus untuk bapak bisa menyalurkan bakat speech bapak.. Aamiin.. Makasih utk mas niko yang membuat video ini :)				
panglima 392	Buat yang ingin kesana,tolong jangan nawar apalagi bilang kemahalan ke bapak ojek.. Karena uang segitu gak sebanding dengan kerugian yang bapak ojek alami. Sukses terus buat chanelnya mas,kita doakan sama sama semoga warga korban lumpur lapindo segera cepat beres uang ganti ruginya.	Positif	Negatif	Positif	Positif
Almecca Ramadhani	Ya allah... Aku terharu liat bapaknya langsung lari ambil motor, semangat banget.. Ramah pula dengan menjelaskan apa saja d.sna... Sehat trus pak... Sing slamet semua nya... Aamiin	Positif	Positif	Negatif	Positif
Nurul Has	Bapak ojek sumringah bgt raut wajahnya pas ada penumpang, sehat sehat bapak lancar rezekinya	Positif	Negatif	Negatif	Negatif
Anang Wahyu	Terharu lihat bpk ojeknya, betapa senengnya dan semangatnya saat ada penumpang yg mau ngojek, smoga di beri rezeki yg banyak pakkk. Amin amin	Positif	Negatif	Negatif	Negatif

B. Implementasi

Implementasi dilakukan dengan mengaplikasikan algoritma analisis sentimen ke dalam sistem. Proses algoritma mencakup tahap preprocessing untuk mengubah dokumen dalam dataset menjadi kumpulan term melalui tokenisasi, pembobotan term menggunakan metode TF-IDF, klasifikasi menggunakan algoritma K-Nearest Neighbor (K-NN), serta tahap evaluasi. Seluruh algoritma tersebut diimplementasikan menggunakan bahasa pemrograman Python.

Ilustrasi dari rancangan Implementasi akan ditunjukkan pada gambar 3.1



C. Pengujian dan Analisis

Pada proses ini akan dijelaskan proses pengujian berdasarkan rancangan yang telah dibuat, serta akan dijelaskan proses analisis dari pengujian yang akan dilakukan.

Kebutuhan software

- Sistem Operasi Windows 11
- Google Chrome Browser
- Microsoft Office 2016
- Microsoft Excel 2016
- Bahasa Pemrograman *Python*
- Library yang digunakan adalah *Sastrawi, Numpy, Pandas, re, string, math, scipy*, dan *taudataNlpTm*

Pembahasan Program

Fungsi utama sistem menginisiasi proses klasifikasi dengan menjalankan program utama yang secara otomatis memanggil fungsi-fungsi pendukung. Library Pandas digunakan untuk membaca file berformat CSV sebagai data latih dan data uji.

Gambar 3.1. Script main program

```
[2] from google.colab import files
      uploaded = files.upload()

# Choose file
data_jumpr.csv
• data_jumpr.csv(text/csv) - 95750 bytes, last modified: 5/12/2025 - 100% done
Saving data_jumpr.csv to data_jumpr.csv

[3] import pandas as pd

df = pd.read_csv("data_jumpr.csv")
df.head()
```

	NAMA	TANGGAL	KOMENTAR	I_PAKAR1	I_PAKAR2	I_PAKAR3	I_FIX
0	Niko_Channel	1 tahun yang lalu	Yang kemaren banyak tanya tentang sejarah dan ...	N	N	N	N
1	Arie J	1 tahun yang lalu	Wah Pak Aksan ini bukan tukang ojek biasa, be...	P	P	N	P
2	panglima 392	1 tahun yang lalu	Buat yang ingin kesana,tolong jangan nawar apa...	P	N	P	P
3	Almecca Ramadhani	1 tahun yang lalu	Ya Allah... in Aku tetahu lat bapaknya tangsu...	P	P	N	P
4	Nurul Has	1 tahun yang lalu	Bapak ojek surringah bgi raut wajahnya pas ada...	P	N	N	N

Next steps: [Generate code with df](#) [View recommended plots](#) [New interactive sheet](#)

```
[3] komentar_list = df['KOMENTAR'].tolist()
      label_list = df['I_FIX'].tolist()

# 0 -> Negatif
# 1 -> Positif
label_fix_list = []
for label in label_list:
    if label == "0":
        label_fix_list.append(0)
    else:
        label_fix_list.append(1)

print(label_fix_list)
```

Tahap berikutnya meliputi perhitungan bobot kata menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF) serta penentuan klasifikasi dengan menerapkan algoritma K-Nearest Neighbor (K-NN).

Gambar 3.2. Script TF-IDF dan K-NN

```
[19] !pip install Sastrawi

Requirement already satisfied: Sastrawi in /usr/local/lib/python3.11/dist-packages (1.0.1)

from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
import re

[21] # Buat stemmer
factory = StemmerFactory()
stemmer = factory.create_stemmer()

[22] def preprocess(text):
    text = text.lower()
    text = re.sub(r'[^a-z\s]', '', text) # hanya huruf dan spasi
    return stemmer.stem(text)

[23] processed_docs = [preprocess(doc) for doc in komentar_list]

# TF-IDF
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(processed_docs)

# Split dan klasifikasi KNN
X_train, X_test, y_train, y_test = train_test_split(X, label_fix_list, test_size=0.7, random_state=42)
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train, y_train)

# Evaluasi
y_pred = knn.predict(X_test)
print(classification_report(y_test, y_pred))
```

Pada hasil penelitian Sebanyak 70% dari keseluruhan data yang digunakan pada tahap test sekaligus validasi, pada uji yang dilakukan pada data test ini menentukan nilai yg di dapat nilai *Precision* Positif (1) 67% dan Negatif (0) 85%, untuk nilai *Recall* Positif 70% dan Negatif 83%, untuk nilai rata rata (f1-score) Positif 69% dan negatif 84% seperti sebagaimana ditunjukkan pada gambar di bawah ini:

Gambar 3.3. Hasil evaluasi pengujian TF-IDF dan K-NN

	precision	recall	f1-score	support
0	0.85	0.83	0.84	242
1	0.67	0.70	0.69	122
accuracy			0.79	364
macro avg	0.76	0.77	0.76	364
weighted avg	0.79	0.79	0.79	364

D. Classification Report Heatmap

Pada fase evaluasi ini, penelitian dilakukan dengan penerapan berbagai metrik evaluasi, seperti peta panas (heatmap), *akurasi*, *presisi*, *recall*, serta *skor f1-score*, guna menilai kinerja model secara komprehensif. Pendekatan ini memberikan pandangan terhadap kemampuan model dalam melakukan klasifikasi secara tepat dan efisien dalam identifikasi klasifikasi. Berikut merupakan gambar Script dan Hasil Evaluasi *Classification Report Heatmap*.

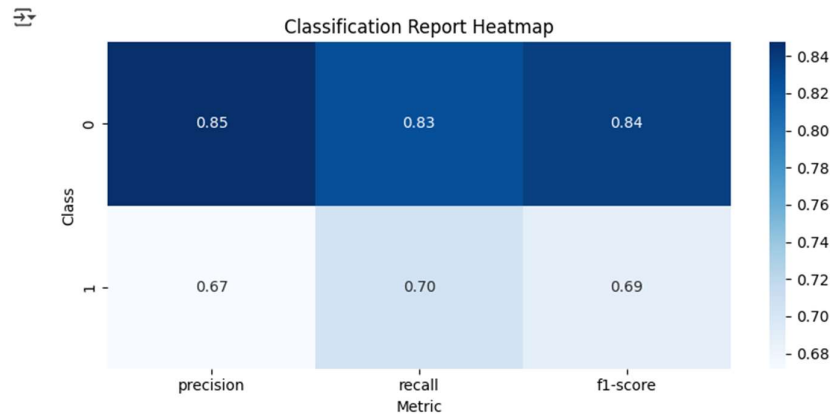
Gambar 3.4. Script Program Report Heatmap

```
# Classification report sebagai dictionary
report_dict = classification_report(y_test, y_pred, output_dict=True)

# Buat DataFrame dan hilangkan 'accuracy', 'macro avg', 'weighted avg'
df_report = pd.DataFrame(report_dict).drop(columns=['accuracy', 'macro avg', 'weighted avg'], errors='ignore').T

# Plot Heatmap
plt.figure(figsize=(8, 4))
sns.heatmap(df_report.iloc[:, :3], annot=True, cmap='Blues', fmt=".2f") # hanya precision, recall, f1-score
plt.title("Classification Report Heatmap")
plt.xlabel("Metric")
plt.ylabel("Class")
plt.tight_layout()
plt.show()
```

Gambar 3.5. Hasil Classification Report Heatmap



Confusion Matrix Heatmap

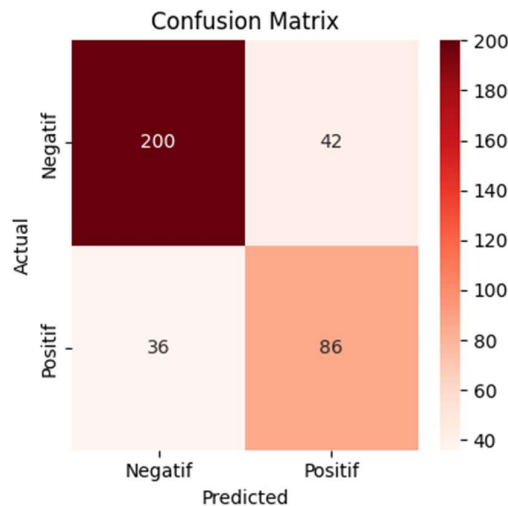
Hasil dari confusion matrix menunjukkan bahwa jumlah True Negative (TN) adalah 200 data, False Positive (FP) sebanyak 42 data, True Positive (TP) sebanyak 86 data, dan False Negative (FN) sebanyak 36 data. Nilai akurasi, presisi, recall, dan f1-score. Kemudian dihitung secara manual berdasarkan data berikut:

Gambar 3.6. Script Program Confusion Matrix Heatmap

```
# Tambahkan Confusion Matrix
conf_matrix = confusion_matrix(y_test, y_pred)

plt.figure(figsize=(4, 4))
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Reds', xticklabels=['Positif', 'Negatif'], yticklabels=['Positif', 'Negatif'])
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix")
plt.tight_layout()
plt.show()
```

Gambar 3.7. Hasil Confusion Matrix Heatmap



Perhitungan nilai manual dari Precision, Accuracy, Recall, f1-score dibawah ini:

$$\text{Accuracy} = \frac{TN+T}{TN+FP+FN+TP} = \frac{200+8}{200+42+36+86} = \frac{286}{364} = 0,78*100 = 78\%$$

$$\text{Precision} = \frac{TP}{TP+F} = \frac{86}{86+42} = \frac{86}{128} = 0,67*100 = 67\%$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{86}{86+36} = \frac{86}{122} = 0,70 \times 100 = 70\%$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0,67 \times 0,70}{0,67 + 0,70} = 68\%$$

IV. SIMPULAN

Dari hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma K-Nearest Neighbor (K-NN) efektif dalam mengklasifikasikan sentimen komentar dengan tingkat akurasi yang memuaskan. Selain itu, penerapan metode TF-IDF sebagai teknik ekstraksi fitur turut berkontribusi signifikan dalam meningkatkan performa model klasifikasi, dengan metode pembobotan **TF-IDF** mampu melakukan klasifikasi sentimen komentar YouTube terhadap video “Pariwisata Lumpur Lapindo” dengan tingkat akurasi sebesar **78%**. Proses pelabelan data dilakukan oleh tiga pakar yang berasal dari Program Studi Psikologi dan menghasilkan label final melalui metode voting. Dari hasil evaluasi terhadap data uji, diperoleh nilai **precision** sebesar **67%**, **recall** sebesar **70%**, dan **f1-score** sebesar **68%** untuk sentimen positif, serta nilai yang lebih tinggi untuk sentimen negatif, menunjukkan bahwa sistem ini cenderung lebih akurat dalam mengenali komentar bernada negatif.

Hasil evaluasi ini menunjukkan bahwa sistem yang dibangun mampu mengidentifikasi sentimen komentar secara cukup baik, meskipun masih terdapat ruang untuk peningkatan akurasi, khususnya pada klasifikasi sentimen positif. Penggunaan algoritma K-NN terbukti efektif untuk tugas klasifikasi berbasis teks seperti ini. Oleh karena itu, sistem ini dapat dijadikan sebagai dasar bagi pengembangan sistem analisis sentimen yang lebih kompleks di masa depan, misalnya dengan menambahkan fitur Bahasa alami (*semantic analysis*), penggunaan algoritma berbasis *machine learning*, atau pengujian terhadap penggunaan dataset yang lebih besar dan beragam diperlukan guna mencapai hasil yang lebih optimal serta meningkatkan kemampuan generalisasi model.

Penulis mengucapkan terima kasih kepada berbagai pihak yang telah memberikan dukungan, baik dalam bentuk tenaga, bimbingan, maupun pemikiran, selama proses penyusunan artikel ini. Ucapan terima kasih secara khusus disampaikan kepada Universitas Muhammadiyah Sidoarjo (UMSIDA) atas fasilitas dan arahan yang diberikan, sehingga penulisan artikel ilmiah ini dapat diselesaikan dengan baik.

Referensi

- [1] A. N. Putra, "Analisis Sentimen Pengguna YouTube Terhadap Uang Baru Tahun Emisi 2022 Menggunakan Metode Naïve Bayes Classifier," *Jurnal Teknologi dan Sistem Komputer*, vol. 5, no. 1, pp. 45–52, Jan. 2024.
- [2] A. Hendrawan dan E. I. Sela, "Analisis Sentimen Komentar Youtube Tentang Resesi Global 2023 Menggunakan LSTM," *J. Indones. Manaj. Inform. dan Komun.*, vol. 5, no. 1, hlm. 587–593, Jan. 2024.
- [3] H. Marlina, E. Elmayati, A. Zulus, and H. O. L. Wijaya, "Penerapan Algoritma Random Forest Dalam Klasifikasi Jurusan Di SMA Negeri Tugumulyo," *Brahmana: Jurnal Penerapan Kecerdasan Buatan*, vol. 4, no. 2, pp. 138–143, Jun. 2023.
- [4] M. H. Mahendra, D. T. Murdiansyah, dan K. M. Lhaksmana, "Analisis Sentimen Tweet COVID-19 menggunakan K-Nearest Neighbors dengan TF-IDF dan Ekstraksi Fitur CountVectorizer," *Dike: Jurnal Ilmu Multidisiplin*, vol. 1, no. 2, pp. 37–43, Agustus 2023.
- [5] M. Djufri, "Penerapan Teknik Web Scraping untuk Penggalan Potensi Pajak (Studi Kasus pada Online Market Place Tokopedia, Shopee dan Bukalapak)," *Jurnal BPPK: Badan Pendidikan dan Pelatihan Keuangan*, vol. 13, no. 2, pp. 65–75, Desember 2020.
- [6] A. E. Budiman dan A. Widjaja, "Analisis Pengaruh Teks Preprocessing Terhadap Deteksi Plagiarisme Pada Dokumen Tugas Akhir," *Jurnal Teknik Informatika dan Sistem Informasi (JuTISI)*, vol. 6, no. 3, pp. 1–8, Des. 2020.
- [7] M. Priandi and Painem, "Analisis Sentimen Masyarakat Terhadap Pembelajaran Daring di Era Pandemi Covid-19 pada Media Sosial Twitter Menggunakan Ekstraksi Fitur CountVectorizer dan Algoritma K-Nearest Neighbor," in *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, Jakarta, Indonesia, Sep. 15, 2021.
- [8] M. A. Rosid, "Improving Text Preprocessing For Student Complaint Document Classification Using Sastrawi," in *Proceedings of the International Conference on Engineering Technology and Social Science (ICETsAS)*, IOP Conf. Series: Materials Science and Engineering, vol. 874, 2020, doi: 10.1088/1757-899X/874/1/012017.

- [9] R. Riyaddulloh dan A. Romadhony, “Normalisasi teks bahasa Indonesia berbasis kamus slang: Studi kasus tweet produk gadget pada Twitter,” *e-Proceeding of Engineering*, vol. 8, no. 4, pp. 42173–42239, Agustus 2021.
- [10] F. Istighfarizky, N. A. S. ER, I. M. Widiartha, L. G. Astuti, I. G. N. A. C. Putra, dan I. K. G. Suhartana, “Klasifikasi jurnal menggunakan metode KNN dengan mengimplementasikan perbandingan seleksi fitur,” *Jurnal Elektronik Ilmu Komputer Udayana (JELIKU)*, vol. 11, no. 1, pp. 167–176, Agustus 2022.
- [11] A. N. H. Regita dan I. Santoso, “Analisis sentimen publik terhadap pengambilalihan jalan rusak di Lampung menggunakan algoritma K-Nearest Neighbors (KNN),” *IKRA-ITH Informatika: Jurnal Komputer dan Informatika*, vol. 7, no. 2, pp. 176–182, Juli 2023.

Conflict of Interest Statement:

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.