Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3485~3492
e-ISSN: 2548-6861

Analysis of User Sentiment Toward Mountain Climbing Content Based on YouTube Comments Using the K-Nearest Neighbors Algorithm

[1] Ghiffari Arrozaq, [2] Sri Mujiono
* Teknik Informatika, Universitas Ngudi Waluyo
** * Teknik Informatika, Universitas Ngudi Waluyo
ghiffariarrozaq@gmail.com[1], mujiyn80@gmail.com [2]

Article Info

Article history:

Received 2026-07-02
Revised 2026-07-24
Accepted 2026-08-08

Keyword:

Sentiment Analysis,
YouTube,
Mountain Climbing,
K-Nearest Neighbor (KNN),
TF-IDF,
Natural Language Processing (NLP).

ABSTRACT

The growth of social media has made YouTube one of the primary sources of information regarding mountain climbing activities. YouTube's comment feature contains a variety of user opinions that can be used to gauge public perception of climbing content. However, the sheer volume of comments makes manual analysis less effective. This study aims to analyze user sentiment toward mountain climbing content based on YouTube comments using the K-Nearest Neighbor (KNN) algorithm. The research method used is a quantitative approach involving data collection through YouTube comment crawling, manual sentiment abelling, and data preprocessing which includes cleaning, case folding, normalization, tokenization, stopword removal, and stemming followed by feature weighting using Term Frequency–Inverse Document Frequency (TF-IDF) and classification using the K -Nearest Neighbor (KNN) algorithm. The model was evaluated using a confusion matrix with the metrics of accuracy, precision, recall, and F1-score. The results of the study show that the K-Nearest Neighbor (KNN) algorithm is capable of classifying the sentiment of YouTube comments with a maximum accuracy of 79.35% at K = 5. These results indicate that the K-Nearest Neighbor (KNN) algorithm is quite effective in analyzing the sentiment of comments related to mountain climbing content, although it still has limitations in understanding semantic context, such as the use of informal language, irony, and sarcasm. Therefore, future research is recommended to use more complex feature representation methods, such as Word2Vec, FastText, or BERT-based models, to improve classification capabilities and sentiment analysis accuracy.

I. INTRODUCTION

Mountain climbing is a form of nature tourism and recreational activity that is currently very popular among Indonesians. This activity is not only practiced by nature lovers or professional climbing communities, but is also popular among the general public who wish to enjoy natural beauty and the experience of climbing mountains. Various mountains in Indonesia have become favorite destinations, frequently visited by thousands of climbers every year. The current high level of public interest in mountaineering indicates that mountain tourism has become an important part of the nature tourism sector in Indonesia. Research on mountaineering shows that people's experiences climbing mountains and exploring the great outdoors can serve as a valuable source of information for evaluating the quality of tourist destinations and the services available [1].

Advances in digital technology have influenced how people obtain information about mountain climbing activities. Social media has become a key source for information such as climbing routes, weather conditions, equipment, and the experiences of previous climbers. Various social media platforms provide a space for users to share their experiences and knowledge about mountaineering, allowing information to spread more quickly and reach a wider audience. This makes social media one of the primary sources of information for climbers before they begin a mountain climb [1].

YouTube is one of the most widely used social media platforms for finding information about mountaineering and the experiences of professional climbers. Users can upload videos documenting their trips, reviews of climbing routes, safety tips, equipment usage, and even personal experiences during their climbs via YouTube. Through the uploaded videos, users can get a clearer picture of on-the-ground conditions than they would from text alone. Therefore, hikers use YouTube as a reference to prepare for their mountain hikes. The growing number of content creators discussing mountain hiking indicates that this topic holds significant appeal among the Indonesian public [1] [2].

YouTube, as an information platform, also offers a comment feature that allows users to share their reactions to uploaded and viewed videos. Comments can take the form of praise, criticism, suggestions, personal experiences, or opinions regarding the information presented in the video. These comments can constitute a large data set containing a variety of public reactions to mountain climbing activities [2]. However, the ever-increasing number of comments has made the manual analysis process less effective and time-consuming.

To address the above issues, an approach capable of processing large volumes of comment data is required. One widely used approach is sentiment analysis. Sentiment analysis is a Natural Language Processing (NLP) technique designed to identify and classify the opinions expressed in comments into positive, negative, or neutral categories. Sentiment analysis has been widely applied to understand public opinion on various social media platforms because it can quickly provide an overview of public sentiment regarding a particular topic.

In the tourism sector, sentiment analysis has been used to evaluate tourists' opinions of various tourist destinations. Hidayat et al. (2026) demonstrated that the K-Nearest Neighbor (KNN) algorithm is capable of delivering strong performance in sentiment analysis of tourism reviews [3]. A study conducted by Latifurrohman (2025) showed that the K-Nearest Neighbor (KNN) algorithm performed well when compared to the Naïve Bayes algorithm in analyzing reviews of the Jepara Ourland Park tourist attraction based on Google Maps [4]. Meanwhile, As'ad (2022) successfully applied the K-Nearest Neighbor (KNN) algorithm to analyze Facebook users' sentiment toward tourist attractions in Central Maluku, yielding fairly good and accurate classification results [5].

Various studies have also shown that the K-Nearest Neighbor (KNN) algorithm performs quite well in sentiment analysis. A study conducted by Larasakti et al. (2023) showed that the K-Nearest Neighbor (KNN) algorithm produced fairly accurate results when analyzing comments on YouTube [6]. Meanwhile, a study by Artamevia and Wibowo (2025) showed that the K-Nearest Neighbor (KNN) algorithm has adequate accuracy in analyzing public sentiment regarding the issue of the Indonesian national team's coaching change [7]. A study conducted by Palepa et al. (2024) showed that the K-Nearest Neighbor (KNN) algorithm is capable of effectively classifying public opinion on political issues [8]. Based on the results of the studies mentioned above, it can be concluded that the K-Nearest Neighbor (KNN) algorithm is one of the most effective algorithms for sentiment analysis.

Although sentiment analysis of YouTube comments has been widely applied in various fields such as politics, the environment, natural disasters, tourism, and many others, a study conducted by Himawan and Sugianto (2024) used the Naïve Bayes and K-Nearest Neighbor (KNN) algorithms to analyze sentiment in videos uploaded to YouTube related to global warming [9]. In their study, Saputra et al. (2026) applied the SVM method to analyze public sentiment toward the flash floods that occurred on the island of Sumatra based on comments on a YouTube video [10]. Another study conducted by Dewati and Sukardani (2024) also utilized YouTube comments to examine users' perceptions of virtual tour content for Bali tourism using the SVM algorithm [11]. The various studies mentioned above demonstrate that YouTube comments can serve as the most effective data source for capturing public opinion. However, research specifically examining user sentiment toward mountain climbing content based on YouTube comments remains very limited, thereby opening opportunities for further research.

Although previous studies have investigated sentiment analysis on YouTube comments in tourism, politics, disasters, and environmental issues, most of them focused on comparing general classification algorithms or evaluating overall model performance. Very limited research specifically investigates user sentiment toward mountain climbing content on YouTube, which has unique linguistic characteristics such as outdoor terminology, climbing jargon, and experience-based expressions. Moreover, previous studies rarely discuss sentiment analysis in the context of mountain tourism content that may provide useful insights for content creators and tourism stakeholders. This limitation indicates a clear research gap that motivates the present study.

On the other hand, the phenomenon of mountain climbing has also garnered public attention on social media. This is evident from a study on public sentiment toward foreign climbers on Mount Rinjani, which shows that the issue of mountain climbing can spark a wide range of public opinions on digital platform [12]. However, research that specifically analyzes user sentiment toward mountain climbing content based on YouTube comments using the K-Nearest Neighbor (KNN) algorithm remains very limited. This situation indicates a research gap that warrants further investigation.

Based on the above description, this study aims to analyze user sentiment toward mountain climbing video content based on YouTube comments using the K-Nearest Neighbor (KNN) algorithm. The results of this study are expected to provide insights into public opinion regarding mountain climbing video content, serve as a basis for evaluation by content creators and mountain climbing tour operators, and enrich the scientific literature on the application of the K-Nearest Neighbor (KNN) algorithm in social media-based sentiment analysis.

II. METHOD

This study uses a quantitative method and applies the K-Nearest Neighbor (KNN) algorithm. In addition, this study aims to determine the performance of the K-Nearest Neighbor (KNN) algorithm in analyzing user sentiment toward mountain climbing content based on YouTube comments. The K-Nearest Neighbor (KNN) algorithm was chosen because many previous studies have shown that it produces fairly good and accurate results in classifying text data [1] [2].

The research process began with the collection of YouTube comment data, sentiment labeling, data preprocessing, feature weighting using TF-IDF, classification using the KNN algorithm, and model evaluation using a confusion matrix. The research workflow is shown in Figure 1.

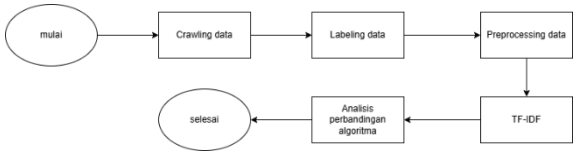


Figure 1. Research Flowchart

Figure 1 shows the workflow of the research method for sentiment analysis of mountain climbing content based on YouTube comments. The study began with data collection from YouTube comments using web crawling techniques, followed by manual data labeling to classify comments as positive or negative. The data was then preprocessed, and a comparative analysis of several algorithms was conducted to identify the algorithm with the highest accuracy.

A. Crawling Data

The data used in this study consists of public comments obtained through a crawling process on the YouTube platform. The data crawling process utilized Google Colab to facilitate the structured collection and processing of large volumes of data. The crawling process was carried out using the YouTube Data API via Python libraries (such as google-api-python-client) running on Google Colab. This process involved retrieving comments based on specific video IDs and filtering the comments based on relevance and text completeness.

Data collection took place in May 2026, yielding 2,000 comments, which were then saved in Excel format (.xlsx) to facilitate processing in the next stage. Following this process, 1,997 comments were deemed suitable for use in the subsequent analysis phase. The reduction of the dataset from 2,000 to 1,997 comments was achieved through data cleaning steps, such as removing duplicates, empty comments, and comments irrelevant to the research topic.

The dataset analyzed in this study consisted of 1,997 YouTube comments collected from mountain climbing-related videos uploaded on YouTube during May 2026. After data cleaning and verification, 1,900 comments were retained for analysis, consisting of 1,120 positive comments (58.95%) and 780 negative comments (41.05%). All comments were written in Indonesian and represented users' opinions regarding mountain climbing content. This dataset was then used as the basis for sentiment classification using the K-Nearest Neighbor algorithm.

B. Labeling Data

At this stage, the data obtained from the crawling process is then labeled or categorized based on the sentiment contained in the comment text. Labeling is performed by dividing the comments into two main categories: positive sentiment and negative sentiment. This labeling stage is crucial in sentiment analysis because it serves as the primary foundation for training the model using the machine learning algorithms employed.

The labeling process is performed manually, taking into account the overall context of the sentences. Each comment is classified into either the positive or negative category based on the opinion expressed. Sentiment labeling was performed manually by one researcher using predefined annotation guidelines. Comments expressing appreciation, satisfaction, or positive experiences were labeled as positive, whereas comments expressing criticism, disappointment, or negative opinions were labeled as negative. To improve labeling consistency, simple annotation guidelines defining the criteria for each sentiment class were used. Since only one annotator was involved, inter-annotator agreement such as Cohen's Kappa could not be calculated and is acknowledged as one limitation of this study. This labeling process helps the system learn specific patterns from the text data, enabling it to automatically classify new data. This stage also aids in understanding public sentiment toward mountain climbing content based on YouTube comments.

C. Preprocessing Data

At this stage, the labeled data undergoes a data preprocessing process. This stage aims to clean and prepare the data so that it is more structured and ready for further analysis. The preprocessing process is crucial in text analysis because raw data typically still contains a significant amount of noise or irrelevant information.

The preprocessing stage in this study consists of several main steps, namely cleaning, normalization, tokenization, stopword removal, and stemming, as shown in Figure 2.

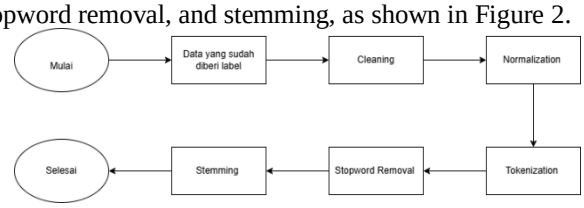


Figure 2. Preprocessing Data

1) Cleaning

The first step in preprocessing is data cleaning, which involves removing irrelevant elements such as symbols, excess punctuation, links (URLs), numbers, user mentions, hashtags, and emojis that have no significant meaning for the analysis. In addition, duplicate comments are also removed to maintain the quality of the dataset.

2) Case Folding

The second step is case folding. This process standardizes the case of letters in a document by converting all uppercase letters to lowercase. This step is performed using the `def lower text` command. The purpose of this process is to avoid differences in meaning caused by the use of capital letters and to ensure consistency in word case before the analysis process.

3) Normalization

The normalization stage in the data preprocessing process involves converting non-standard words, abbreviations, and regionally specific colloquial language into a more organized, consistent format that is ready for analysis using machine learning algorithms. This stage is very important because the comment sections on YouTube frequently contain informal language.

4) Tokenization

Tokenization is the process of breaking text down into small units called tokens, such as words or characters. Tokenization aims to simplify the feature extraction process, in which each token is treated as a potential feature in weight calculations using the TF-IDF method.

5) Stopword Removal

The stopword removal stage is performed to remove common words that are not relevant to the analysis process; if they are not removed, they will interfere with the analysis process and affect the classification results. Examples include the words “apa,” “di,” or “untuk”.

6) Stemming

Stemming is the process of converting a word to its base form by removing affixes, such as prefixes or suffixes. The stemming process is performed using the Sastrawi library, a popular tool in Indonesian research for stemming and stopword removal in local languages. Using this tool has been shown to improve the accuracy of classification systems.

D. TF-IDF

After undergoing text processing, the text data must be converted into numerical values so that it can be easily analyzed by the K-Nearest Neighbor (KNN) algorithm. TF-IDF (Term Frequency-Inverse Document Frequency) is used to identify the most important or frequently occurring words in a dataset or document, while distinguishing them from less relevant words. Each unique word in the document is assigned a weight based on the TF-IDF calculation.

Each unique word will be assigned a weight based on the TF-IDF calculation, which is mathematically formulated to represent the importance of a word in the context of sentiment analysis, as follows.

$$TF - IDF(t,d) = TF(t,d) \times IDF(t)$$

With:

$$IDF(t) = \log\left(\frac{N}{df(t)}\right)$$

Notes:

- TF (t,d) : frequency of occurrence of term t in document d
- Df (t) : number of documents containing term t
- N : total number of documents

E. K-Nearest Neighboors Algorithm

The K-Nearest Neighbor (KNN) algorithm is one of the methods used for classification. The K-Nearest Neighbor (K-NN) algorithm works by comparing the test data to the training data points that are closest to it. In this study, the K-Nearest Neighbor (KNN) algorithm was used to classify the sentiment of YouTube comments into positive and negative categories based on text feature representations weighted by TF-IDF.

The selection of the K-Nearest Neighbor (KNN) algorithm was based on its simple concept, an easily understandable classification process, and its ability to handle high-dimensional data. Furthermore, KNN does not require complex model training, making it highly suitable for sentiment analysis on social media, where informal language is commonly used. In this study, distances were calculated using the Euclidean distance method because this technique is widely used to measure the similarity between numerical vectors and is well-suited to the characteristics of TF-IDF features.

K-Nearest Neighbor was chosen because this method provides a practical classification approach, performs very well on sparse TF-IDF vectors, does not require complex model training, and has demonstrated competitive performance on medium-sized sentiment datasets. Although algorithms such as Support Vector Machines, Naïve Bayes, Random Forests, and Transformer-based models often achieve higher classification accuracy, K-Nearest Neighbor was chosen in this study as a computationally efficient baseline. Given the objective of this study to evaluate sentiment classification in mountain climbing comments using a classical machine learning approach K-Nearest Neighbor offers a good balance between simplicity, readability, and computational efficiency.

The steps for applying the K-Nearest Neighbor (KNN) method in this study are described as follows.

1) Determining the Value of K
   Determine the number of nearest neighbors that will be used in the classification process, while also finding the optimal value of K through cross-validation. Generally, K is chosen as an odd number to avoid balanced classification results.
2) Calculating Distance
   Measuring the distance between training data and test data using the Euclidean distance method ensures that the process of comparing a new sentence with all sentences in the training data runs smoothly.
3) Determining the Nearest Neighbors
   Based on the distance calculations, select K data points with the shortest distances. These data points are considered the most relevant to the new data to be analyzed. This data will later be used as the basis for determining sentiment.
4) Determining the Class
   Determine the category based on the class that appears most frequently among the K nearest neighbors, using a majority voting process to classify the new data.

Although KNN has advantages in terms of simplicity and ease of interpretation, the KNN algorithm also has limitations in handling informal text in a semantic context, such as sarcasm and irony, which frequently appear in comments on the TikTok app. These limitations are one of the factors affecting the algorithm's accuracy, which is not yet optimal; therefore, the classification results need to be analyzed further in the discussion chapter.

F. Evaluation

At this stage, the results of applying the K-Nearest Neighbor (KNN) classification algorithm were evaluated to determine how well the model could accurately and precisely classify sentiment toward mountain climbing content based on YouTube comments. The evaluation presents accuracy, precision, recall, and F1-score to help demonstrate the model's accuracy. This evaluation approach aims to provide a more comprehensive picture of the model's performance, particularly in handling the characteristics of social media sentiment data, which often exhibits class distribution imbalances.

In this evaluation process, the first step is to determine and calculate the number of true positives, false positives, false negatives, and true negatives for each row in the comment dataset. These values will be used to calculate the model's performance metrics, such as accuracy, precision, recall, and F1-score.

III. Results And Discussion

This section presents the results of data processing obtained from YouTube using the K-Nearest Neighbor (KNN) algorithm. The results are presented in stages, beginning with data crawling, data labeling, data preprocessing, and algorithm evaluation. The discussion is structured to explain the relationship between the research findings and existing theories as well as the results of previous studies, thereby providing a clear picture of the algorithm's ability to analyze public sentiment.

A. Data Collection

This study used public comments collected from YouTube videos related to mountain climbing content. The comments were retrieved using the YouTube Data API through Google Colab. Initially, a total of 2,000 comments were collected. After the data cleaning process, which involved removing duplicate comments, empty comments, and comments irrelevant to the research topic, 1,997 comments remained and were used in the subsequent sentiment analysis.

The collected comments contain various public opinions regarding mountain climbing content, including positive and negative sentiments. These comments serve as the primary dataset for evaluating the performance of the K-Nearest Neighbor (KNN) algorithm in sentiment classification. Figure 3 presents the results of the data collection process.

	publishedAt	authorDisplayName	textDisplay	likeCount
0	2020-11-01T11:35:10Z	@TheSlackerHiker	Sudahkah anda ke Cemoro Kandang?	335
1	2020-11-01T11:38:44Z	@ganongadventure3306	Wingi tektok nk kawah mbak	8
2	2020-11-01T11:45:26Z	@danangariyanto2754	Sudah	3
3	2020-11-01T11:45:48Z	@danangariyanto2754	Nyoba jalur jogorogo Ngawi kak. Biar gw jalurY	4
4	2020-11-01T11:56:28Z	@cahyoagungsutrisno3414	Uwes mbak soale mbahku asli magetan, tapi duru...	4
...	...	...	...	...
1993	2020-11-01T09:01:31Z	@mohammadsholehuluwwi8705	Akhirnya update lagi mbak uki	0
1994	2020-11-01T09:01:21Z	@rikimuhamadikhlas7256	1 jam Gas	2
1995	2020-11-01T09:01:11Z	@kautsar4821	Ini yg di tungu2 bidadari medok	4
1996	2020-11-01T09:01:10Z	@bodoamatstory	52 detik yang lalu	1
1997	2020-11-01T09:01:05Z	@manuremvo	Pertama heheeeee	2

1998 rows × 4 columns

Figure 3. Data Collection Results

B. Labeling Data

After the comment crawling process, the next step was manual sentiment labeling. Each comment was categorized into either a positive or negative sentiment class based on its contextual meaning. The labeling process was performed manually by the researcher using predefined labeling guidelines to ensure consistency. A total of 1,997 comments were successfully labeled and used in the sentiment classification process. Table I presents several examples of the labeled data.

TABLE I
LABELING DATA

Label	Text
Negative	Mungkin kemaren cuma papasan, aku Jumat jam naiknya
Positive	keren banget pembuatan vidionya, sinematik habis

The following visualization compares the number of positive and negative sentiments shown in Figure 3, indicating that the proportion of negative sentiments is much higher.

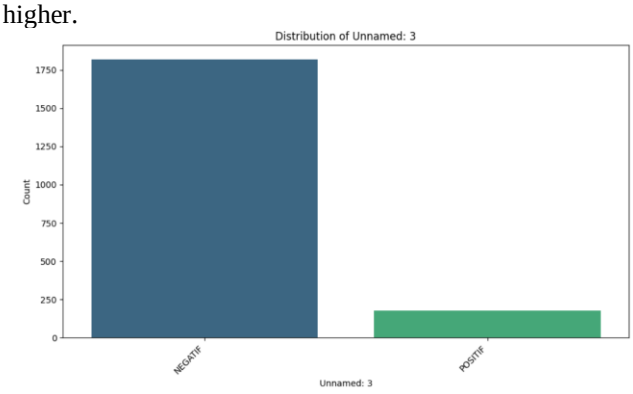


Figure 4. Comparison of positive and negative sentiment

C. Preprocessing Data

The data processing stage is conducted to evaluate the quality of comment data from the YouTube app before it is used in sentiment analysis. Given that comments on the YouTube app are generally informal and contain abbreviations, spelling variations, and non-text elements, this stage plays a crucial role in reducing noise and improving the quality of the text representation. The results of each text processing stage not only reveal changes in the data format but also provide a solid foundation for TF-IDF weighting and

the performance of the K-Nearest Neighbor (KNN) algorithm in classifying the data.

1) Cleaning

The initial stage of data processing involves data cleaning to ensure that the comments on the YouTube app to be analyzed truly represent public opinion in textual form. The raw data obtained from crawling still contains many URLs, emojis, numbers, symbols, duplicates, and empty fields that do not contribute to sentiment analysis. The presence of these non-text elements can interfere with the analysis process.

An example of the text after the data cleaning process is presented in Table II, where the comments have been stripped of symbols and non-text elements, making them easier to process in the next stage.

TABLE II
CLEANING DATA

Label	Text
Negative	Mungkin kemaren cuma papasan, aku Jumat jam naiknya
Positive	keren banget pembuatan vidionya, sinematik habis

2) Case Folding

Once the data cleaning stage is complete, the data will be processed using case folding. Case folding is the process of converting all uppercase letters in the comment text to lowercase. The purpose of this stage is to eliminate differences in word representation caused by variations in the use of uppercase and lowercase letters. Without case folding, a word written in uppercase would have the same meaning as its lowercase counterpart, which the system would then treat as two distinct features.

The results of applying case folding are shown in Table III, which demonstrates that all comments have been processed using case folding. This process reduces feature redundancy and improves the consistency of text representation prior to feature weighting..

TABLE III
CASE FOLDING

Label	Text
Negative	mungkin kemaren cuma papasan, aku jumat jam naiknya
Positive	keren banget pembuatan vidionya, sinematik habis

3) Normalization

The normalization stage is performed to convert non-standard words, abbreviations, and excessive spaces into a more standard form of language. Language on social media is very informal, often characterized by the use of abbreviations such as “yg,” “sdh,” or “udh”; if normalization is not performed, this will result in the fragmentation of features with the same meaning.

The results of the normalization can be seen in Table VI below, where the abbreviations have been converted to their standard forms. This normalization helps improve text comprehension and ensures that words with similar meanings are represented consistently in the sentiment analysis process.

TABLE VI
NORMALISASI

Label	Text
Negative	mungkin kemaren cuma papasan, aku jumat jam naiknya
Positive	keren banget pembuatan vidionya, sinematik habis

4) Tokenization

Next, the tokenization process is performed to break the text down into individual words (tokens). Through this process, each word can be recognized as an independent feature that is subsequently analyzed based on its frequency of occurrence. Tokenization also plays a crucial role in classification using the K-Nearest Neighbor (KNN) algorithm by utilizing feature representations in the form of word vectors.

The results of the tokenization can be seen in Table V, which shows that each comment sentence has been broken down. With this step completed, weight calculations can be performed accurately.

TABLE V
TOKENIZATION

Label	Text
Negative	mungkin, kemaren, cuma, papasan, aku, jumat, jam, naiknya
Positive	keren, banget, pembuatan, vidionya, sinematik, habis

5) Stopword Removal

The next step is stopword removal, a process that removes common words that have no bearing on sentiment analysis. Words that frequently appear in the text include “and,” “at,” and “to.” The results of the stopword removal are shown in Table VI, which displays the common words that have been removed, leaving only the informative words.

TABLE VI
STOPWORD REMOVAL

Label	Text
Negative	mungkin, kemaren, cuma, papasan, aku, jumat, jam, naik
Positive	keren, banget, pembuatan, vidio, sinematik, habis

6) Stemming

Stemming is the final stage of text processing in this study. The stemming process is performed to convert words into their base forms by removing prefixes, suffixes, or infixes. In this study, stemming was performed using the Sastrawi library, which was specifically designed for the Indonesian language and has been widely used in sentiment analysis research by previous researchers.

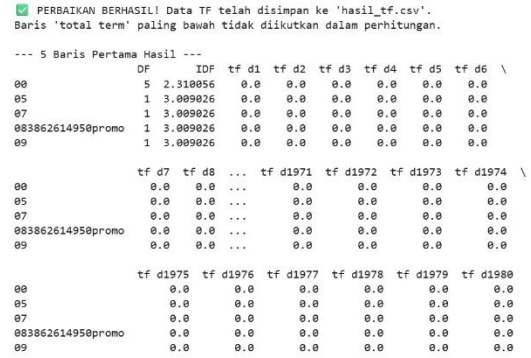
Table VII shows the results of the stemming process, with words simplified to their base forms. This process aims to consolidate variations of words that share the same meaning, thereby improving the consistency of features in the sentiment classification process.

TABLE VII
STEMMING

Label	Text
Negative	Mungkin kemaren cuma papasan aku jumat jam naik
Positive	keren banget pembuatan vidio sinematik habis

D. TF-IDF

TF-IDF is used to measure how important a word is in a document relative to all documents. The higher a word's TF-IDF value, the greater its influence or weight in representing the document's content. TF-IDF is a commonly used method in text processing because it assigns a weight to each word based on its level of importance within a document relative to all existing documents.



```
✅ PERBAIKAN BERHASIL! Data TF telah disimpan ke 'hasil_tf.csv'.
Baris 'total term' paling bawah tidak diikutkan dalam perhitungan.

--- 5 Baris Pertama Hasil ---
                    DF       IDF  tf d1  tf d2  tf d3  tf d4  tf d5  tf d6  \
00                   5  2.310056    0.0    0.0    0.0    0.0    0.0    0.0
05                   1  3.009026    0.0    0.0    0.0    0.0    0.0    0.0
07                   1  3.009026    0.0    0.0    0.0    0.0    0.0    0.0
083862614950promo    1  3.009026    0.0    0.0    0.0    0.0    0.0    0.0
09                   1  3.009026    0.0    0.0    0.0    0.0    0.0    0.0

                    tf d7  tf d8  ...  tf d1971  tf d1972  tf d1973  tf d1974  \
00                    0.0    0.0  ...       0.0       0.0       0.0       0.0
05                    0.0    0.0  ...       0.0       0.0       0.0       0.0
07                    0.0    0.0  ...       0.0       0.0       0.0       0.0
083862614950promo     0.0    0.0  ...       0.0       0.0       0.0       0.0
09                    0.0    0.0  ...       0.0       0.0       0.0       0.0

                    tf d1975  tf d1976  tf d1977  tf d1978  tf d1979  tf d1980
00                       0.0       0.0       0.0       0.0       0.0       0.0
05                       0.0       0.0       0.0       0.0       0.0       0.0
07                       0.0       0.0       0.0       0.0       0.0       0.0
083862614950promo        0.0       0.0       0.0       0.0       0.0       0.0
09                       0.0       0.0       0.0       0.0       0.0       0.0
```

Figure 5. TF-IDF

E. K-Nearest Neighbors Algorithm

At this stage, the sentiment classification process is performed using the K-Nearest Neighbors (KNN) algorithm to evaluate the model's ability to classify text into positive and negative classes. The dataset from the previous stage is split into 90% training data and 20% test data using the `train_test_split` function. This division aims to ensure that the model is trained with sufficiently representative data and tested using data that has not been encountered before. The next step is to determine the optimal value of K. Based on testing values of K ranging from 1 to 15, the highest accuracy was achieved at K = 4, with an accuracy of 70.50% using the Euclidean metric. The decline in performance for K values above 4 indicates that the more neighbors involved in the classification process, the greater the likelihood of irrelevant data being included in the majority voting process. This can cause the model to become less sensitive to locally patterned data. This condition is quite common in KNN algorithms, especially when applied to high-dimensional text data such as TF-IDF representations.

The value of K was also further evaluated by comparing odd-numbered values of K from 1 to 15. In this experimental setting, K = 5 produced the highest evaluation performance while maintaining a balance between bias and variance. Smaller K values tended to be more sensitive to noisy data, whereas larger K values reduced the influence of local neighborhood information. Based on this analysis, K = 5 was selected as the final parameter used in the model evaluation stage due to its more stable performance across different test splits.

In addition, the moderate accuracy rate may also be influenced by the nature of comments on the YouTube app, which are generally short, informal, and rich in linguistic variations such as puns, abbreviations, and sarcasm. This makes it difficult for distance-based algorithms to comprehensively understand sentiment. It is not uncommon for comments that explicitly appear neutral or positive to actually imply a negative meaning, thereby increasing the potential for errors during the classification process. Therefore, this study demonstrates that while KNN can serve as an initial approach for sentiment classification on social media, this method still has significant limitations in processing the natural language used on platforms like YouTube. The findings are also consistent with previous research indicating that classical algorithms such as K-Nearest Neighbor (KNN) exhibit suboptimal performance when applied to unstructured social media data.

F. Evaluation

Based on the results of testing the sentiment classification model using the K-Nearest Neighbor algorithm, the best accuracy rate of 79.35% was achieved at K = 5. These results indicate that the K-Nearest Neighbor algorithm is capable of classifying YouTube user comments regarding mountain climbing content with a fairly high degree of accuracy. The evaluation results show that the application of preprocessing steps including cleaning, case folding, normalization, tokenization, stopword removal, and stemming has a significant impact on improving data quality before the classification process is carried out. These steps reduce the presence of irrelevant words, standardize word forms, and produce a better data representation for the weighting process using the TF-IDF method. Consequently, the K-Nearest Neighbor algorithm can calculate the similarity between comments more accurately.

To provide a clearer position of the proposed method compared to previous studies, a comparison with related research is presented. Larasakti (2023) applied the K-Nearest Neighbor (KNN) algorithm for YouTube comment sentiment analysis and obtained an accuracy of approximately 78%. Meanwhile, Saputra et al. (2026) used the Support Vector Machine (SVM) method for sentiment analysis of flood-related YouTube comments and achieved an accuracy of around 83%. In another study, Dewati and Sukardani (2024) also applied SVM for analyzing YouTube comments on virtual tourism content and reported an accuracy of 81%. Compared to these studies, the proposed model in this research using KNN achieved an accuracy of 79.35%.

Compared with previous studies, the proposed model achieved comparable performance. Although several SVM-based studies reported higher accuracy, the obtained results indicate that the K-Nearest Neighbor (KNN) algorithm remains a competitive and computationally efficient approach for sentiment analysis, particularly for YouTube comments related to mountain climbing content.

The results of this study indicate that the K-Nearest Neighbor algorithm is quite effective for sentiment analysis of YouTube comments regarding mountain climbing content. These results are also consistent with several previous studies that suggest the K-Nearest Neighbor algorithm performs well in TF-IDF-based text classification, although it still has limitations in understanding the semantic context of natural language. Therefore, future research may consider using more complex feature representation methods, such as Word2Vec, FastText, or BERT-based models, to improve the system's ability to understand linguistic context and enhance classification accuracy.

IV. CONCLUSION

Based on the results of a study analyzing user sentiment toward mountain climbing content based on YouTube comments using the K-Nearest Neighbor (KNN) algorithm, it can be concluded that the KNN algorithm can be used to classify user comments into positive and negative sentiment categories with a fairly good level of performance. The analysis process involved collecting YouTube comments, manually labeling sentiment, preprocessing the data—which included cleaning, case folding, normalization, tokenization, stopword removal, and stemming—and then weighting features using the TF-IDF method before performing classification using the KNN algorithm.

The evaluation results show that implementing the preprocessing steps improves data quality, thereby optimizing the classification process. Based on model testing, the KNN algorithm achieved the best accuracy of 79.35% at K = 5, indicating that this method is quite effective in performing sentiment analysis on YouTube comments regarding mountain climbing content.

Nevertheless, this study also shows that the KNN algorithm still has limitations in understanding the semantic context of natural language, such as the use of informal language, abbreviations, irony, and sarcasm, which are commonly found in social media comments. Therefore, future research is recommended to utilize more complex feature representation methods, such as Word2Vec, FastText, or BERT-based models, so that the system's ability to understand linguistic context can be improved and result in better classification accuracy.

Based on the dataset used in this study, the KNN algorithm demonstrated satisfactory performance for sentiment classification of YouTube comments related to mountain climbing content. Therefore, the findings should be interpreted within the scope of this dataset and cannot yet be generalized to all social media sentiment analysis tasks.

The findings may assist YouTube content creators in understanding audience responses, help national park managers evaluate public perceptions regarding mountain climbing activities, and support tourism stakeholders in improving educational content related to mountain safety and environmental conservation.

REFERENCES

[1] Hidayat, A. N., Ramdani, A. L., & Muthoharoh, L. (2026). Analisis Sentimen Berbasis Aspek pada Ulasan Pariwisata Menggunakan Varian Algoritma K-Nearest Neighbor: Aspect-Based Sentiment Analysis of Tourism Reviews Using Variants of the K-Nearest Neighbor Algorithm. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 6(2), 534-545. https://www.journal.irpi.or.id/index.php/malcom/article/view/2522

[2] Latifurrohman, L. (2025). Perbandingan Perbandingan Metode K-Nearest Neighbor (Knn) Dan Naive Bayes Untuk Analisis Sentimen Review Objek Wisata Jepara Ourland Park (JOP) Berdasarkan Ulasan Pada Google Maps. Jtinfo: Jurnal Teknik Informatika, 4(2), 277-286. https://journal.unisnu.ac.id/JTINFO/article/view/1296

[3] As' ad, I. (2022). Analisis Pengguna Facebook Terhadap Objek Wisata Di Maluku Tengah Menggunakan Metode K-Nearest Neighbor. Buletin Sistem Informasi dan Teknologi Islam. https://jurnal.fikom.umi.ac.id/index.php/BUSITI/article/view/1329

[4] Artamevia, A. M., & Wibowo, J. S. (2025). Analisis Sentimen Terhadap Kontroversi Pergantian Pelatih Timnas Indonesia Menggunakan Metode Knn (K-Nearest Neighbour). JATI (Jurnal Mahasiswa Teknik Informatika), 9(5), 8945-8952. https://www.ejournal.itn.ac.id/jati/article/view/15228

[5] Palepa, M. J., Pratiwi, N., & Rohmansa, R. Q. (2024). Analisis Sentimen Masyarakat Tentang Pengaruh Politik Identitas Pada Pemilu 2024 Terhadap Toleransi Beragama Menggunakan Metode K-Nearest Neighbor. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 9(1), 389-401. https://jurnal.stkippgritulungagung.ac.id/index.php/jipi/article/view/4957

[6] Himawan, A., & Sugianto, C. A. (2024). Perbandingan algoritma Naïve Bayes dan K-Nearest Neighbor (KNN) pada video YouTube mengenai global warming. Informatics and Digital Expert (INDEX), 6(2), 98-104. https://e-journal.unper.ac.id/index.php/informatics/article/view/1803

[7] Larasakti, D. N., Aziz, A., & Aditya, D. (2023). Analisis Sentimen Komentar Video Youtube Dengan Metode K-Nearest Neighbor. Jurnal Ilmiah Wahana Pendidikan, 9(5), 132-142. https://jurnal.peneliti.net/index.php/JIWP/article/view/3953

[8] Saputra, A., Nurdiyani, I., Nurhidayah, U. S., & Maesaroh, S. (2026). Analisis Sentimen Masyarakat Terhadap Penanganan Banjir Bandang di Pulau Sumatra Berdasarkan Komentar Youtube Menggunakan Metode Support Vector Machine (SVM). Jejak digital: Jurnal Ilmiah Multidisiplin, 2(1), 2225-2239. http://indojurnal.com/index.php/jejakdigital/article/view/2094

[9] Dewati, P. P. P., & Sukardani, P. S. (2024). Analisis Sentimen Komentar Youtube Virtual Tour 360° Wisata Bali Oleh Kemenparekraf. The Commercium, 8(2), 146-153. https://ejournal.unesa.ac.id/index.php/Commercium/article/view/62511

[10] Suta, P. W. P., Rafael, R., & Kesumadewi, A. A. A. R. (2025). Analisis Ulasan Wisatawan Pada Wisata Trekking Di Batur. Jurnal Ilmiah Hospitality, 14(2), 719-728. https://ejournal.stpmataram.ac.id/JIH/article/view/4012

[11] Hadi, M. S., Akbar, J., & Zulkarnain, M. F. (2025). Analisis Sentimen Wisata Air Terjun Di Kabupaten Lombok Tengah Menggunakan Metode Support Vector Machine (SVM). SKANIKA: Sistem Komputer dan Teknik Informatika, 8(2), 318-329. https://jom.fti.budiluhur.ac.id/SKANIKA/article/view/3578

[12] Satio, D. A., Fallah, A. H. N., Setiawan, A., & Kuncoroyakti, Y. A. (2026). Analisis Isi Komentar pada Media Sosial YouTube Fiersa Besari" Pendaki Fomo?". Jurnal Penelitian Ilmiah Multidisipliner, 2(03), 2226-2233. https://ojs.ruangpublikasi.com/index.php/jpim/article/view/1143

[13] Novitasari, D. C., & Putri, R. R. (2025). Analisis Sentimen Cuitan X terhadap Pendaki Asing Gunung Rinjani Menggunakan Algoritma SVM dan Random Forest. INTEGER: Journal of Information Technology, 10(2). https://ejurnal.itats.ac.id/integer/article/view/8119