



International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Sentiment Analysis of Social Media Presence by using Distil Bert

Tejas M, Ramanujam DK, Akash S, Bhuvan Cariappa BD, Marimuthu K

Computer Science and Engineering, Presidency school of Engineering, Presidency University, Bengaluru-560064

ABSTRACT :

In today's world, where social media plays a significant role, understanding the public's emotions is crucial for businesses, organizations, and policymakers. This study focuses on analyzing the sentiments expressed in social media posts by using a blend of advanced methods, including data collection, text preprocessing, and embedding strategies. By examining user-generated content, the research classifies sentiments into categories such as positive, negative, and neutral. It evaluates the analysis through various metrics like accuracy, precision, and sentiment trends, offering a detailed view of emotional expression. Furthermore, the study tracks how sentiments change over time, identifying shifts and emerging patterns. It also explores how both text and visual elements contribute to sentiment, offering deeper insights into online communication. The study links user engagement indicators, such as likes, shares, and comments, with sentiment types, highlighting the effect of emotions on social media behaviour. By combining quantitative data with qualitative observations, the research uncovers challenges in sentiment analysis and continuously improves the model through user input. This analysis provides businesses and policymakers with actionable insights to better understand and react to public sentiment.

Keywords: Social media platform, Sentiment analysis, Twitter, Instagram, Natural Language Processing, DistilBERT

INTRODUCTION:

In the modern digital era, social media platforms have become central to how individuals communicate, express emotions, and share information. With billions of users across platforms like Twitter, Facebook, Instagram, and YouTube, these digital spaces generate vast amounts of content that offer rich insights into public sentiment and societal trends. Understanding these sentiments is crucial for businesses, policymakers, and individuals who wish to stay informed and make data-driven decisions. This is where sentiment analysis, a key branch of natural language processing (NLP), plays a vital role in deciphering the emotions, opinions, and attitudes expressed in both text and visual content. The challenge lies in accurately interpreting complex human emotions within diverse, informal, and often nuanced language used on social media. Moreover, with the integration of machine learning and advanced computational techniques, sentiment analysis has become more sophisticated, allowing deeper insights into public opinion. This research explores the intricacies of sentiment analysis within social media, examining the emotional landscape of posts, comments, and interactions across platforms, and providing valuable tools to navigate the ever-evolving digital environment.

i Background:

The proliferation of social media platforms like Twitter, Facebook, Instagram, and YouTube has transformed the way individuals communicate, express emotions, and share information. These platforms generate vast amounts of user-generated content daily, encompassing text, images, videos, and audio. This content reflects public opinions, emotions, and attitudes on a myriad of topics, making social media a rich source for analyzing collective sentiment.

Sentiment Analysis (SA), a subfield of Natural Language Processing (NLP), focuses on computationally identifying and categorizing opinions expressed in text and other media formats. Traditional SA techniques primarily dealt with textual data; however, the emergence of multimodal content necessitates more sophisticated approaches that can handle diverse data types, including images and videos.

ii Research Motivation:

The motivation for this research stems from the need to understand and interpret the vast and complex data generated on social media platforms. Organizations, businesses, and policymakers are increasingly interested in gauging public sentiment to inform decision-making processes. However, the informal, diverse, and often ambiguous nature of social media content presents significant challenges for accurate sentiment analysis.

Moreover, the integration of multimodal data—combining text with images, audio, and video—adds layers of complexity that traditional SA methods are ill-equipped to handle. Addressing these challenges requires the development of advanced analytical tools and methodologies capable of processing and interpreting multimodal content effectively.

iii Objectives:

The primary objective of this research is to design and implement a robust sentiment analysis framework capable of handling the diverse and dynamic nature of social media content. By integrating advanced machine learning and deep learning algorithms, the framework aims to accurately interpret sentiments expressed through both textual and visual data. It seeks to enable real-time monitoring of public sentiment, uncover emotional trends, and enhance the overall understanding of user-generated content. This approach not only improves analytical precision but also supports data-driven decision-making for businesses, policymakers, and researchers, ultimately offering deeper insights into public opinion across digital communication platforms.

iv Significance of Research:

This research holds significant value in multiple dimensions:

- **Enhanced Decision-Making:** By providing accurate and timely insights into public sentiment, organizations and policymakers can make more informed decisions.
- **Improved Customer Engagement:** Businesses can better understand customer needs and preferences, leading to improved products, services, and customer satisfaction.
- **Advancement of Academic Knowledge:** The research contributes to the academic field by addressing existing gaps in multimodal sentiment analysis and proposing novel methodologies.
- **Societal Impact:** Understanding public sentiment on critical issues can aid in addressing societal challenges, shaping public policy, and fostering community engagement.

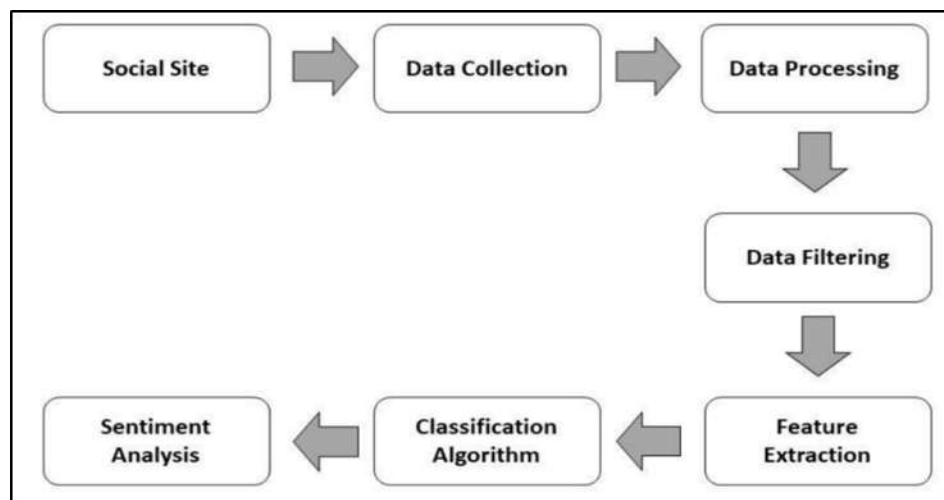


Figure 1: System architecture of sentiment analysis

RELATED WORKS:

For this study, we selected **DistilBERT (distilbert-base-uncased)** as the foundational model due to its efficiency and performance. DistilBERT is a distilled version of BERT, retaining 97% of BERT's language understanding capabilities while being 60% faster and 40% smaller. This makes it particularly suitable for real-time sentiment analysis tasks on social media data. The model architecture was extended by adding a classification head—a fully connected dense layer—on top of the pre-trained DistilBERT base to facilitate ordinal multiclass classification, predicting sentiment scores ranging from 1 to 5. This setup allows the model to capture nuanced sentiment variations in user-generated content. The implementation leveraged the Hugging Face Transformers library, which provides a streamlined interface for model customization and deployment.

Pros:

Sentiment analysis on social media platforms offers the significant advantage of processing vast amounts of user-generated content in real-time, providing timely and actionable insights into public opinion. This ability is highly beneficial for businesses, policymakers, and researchers who aim to monitor public sentiment and respond strategically. Moreover, the integration of advanced machine learning and deep learning techniques has substantially improved the accuracy of sentiment classification, especially when dealing with large and diverse datasets. Multimodal sentiment analysis, which incorporates textual, visual, and sometimes audio data, further enriches the understanding of user emotions and behaviours on social media platforms.

Cons:

Despite its strengths, sentiment analysis on social media faces several challenges. Accurately interpreting sentiments expressed through sarcasm, irony, or culturally nuanced language remains a significant hurdle. The informal, unstructured nature of social media content — characterized by slang, abbreviations, misspellings, and emojis — complicates traditional NLP processing. Additionally, the ever-evolving language trends on social platforms require models to constantly adapt, making sustained accuracy difficult. Although multimodal analysis offers deeper insights, it also introduces

complexities in data alignment, feature extraction, and model training, demanding more computational resources and sophisticated algorithms to maintain effectiveness.

MATERIALS AND METHODS:

DistilBERT is a compact, faster version of BERT created through knowledge distillation. It retains most of BERT's performance while reducing size and training time. In sentiment analysis, it is fine-tuned on labelled data using a classification layer, enabling efficient and accurate prediction of sentiment from textual input.

Dataset Preparation and Tokenization

The Amazon Fine Food Reviews dataset, comprises over 500,000 reviews of fine foods from Amazon, collected over a span of more than 10 years up to October 2012. Each entry includes attributes such as Product ID, User ID, Profile Name, Helpfulness rating, Score (ranging from 1 to 5), Time, Summary, and the full Text of the review. This rich dataset is widely used for sentiment analysis tasks, allowing researchers to explore customer opinions, preferences, and sentiments toward various food products. The dataset's extensive size and diversity make it suitable for training and evaluating machine learning models aimed at understanding consumer behaviour and improving recommendation systems.

Training and Fine-tuning process

The fine-tuning process utilized the API, which simplifies the training loop and integrates seamlessly with the Transformers library. The model was trained using the AdamW optimizer with a learning rate of $5e-5$ and a batch size of 32 over multiple epochs. During training, the classification head's weights (classifier.weight and classifier.bias) and the pre-classifier layer's weights (pre_classifier.weight and pre_classifier.bias) were initialized and updated, tailoring the model to the specific sentiment analysis task. The training process included evaluation on the test set after each epoch to monitor performance metrics such as accuracy, precision, recall, and F1-score. This iterative approach ensured the model's robustness and generalizability to unseen social media data.

Model Architecture

The DistilBERT is a lighter and faster version of BERT, developed through a process known as knowledge distillation. In this process, a smaller "student" model is trained to replicate the behaviour of a larger "teacher" model (BERT), capturing its key knowledge while reducing computational complexity. Architecturally, DistilBERT cuts the number of transformer layers from 12 in BERT to 6, resulting in a model that is about 40% smaller and runs 60% faster, with minimal loss in accuracy—retaining approximately 97% of BERT's performance. It also removes components such as token-type embeddings and the pooler, streamlining the model further. DistilBERT is trained using a combination of three loss functions: masked language modelling loss, distillation loss, and cosine embedding loss, which help the student model generalize effectively while preserving the linguistic richness of its predecessor. This architecture makes DistilBERT highly efficient for real-time applications like sentiment analysis, especially when computational resources are limited.

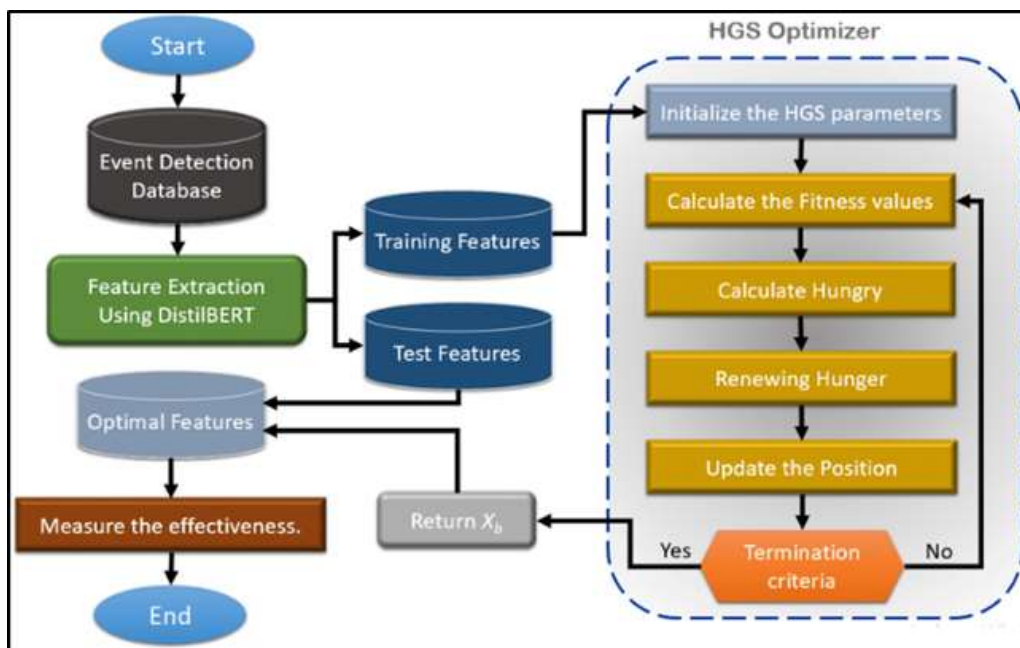


Figure 2: DistilBERT Flow chart

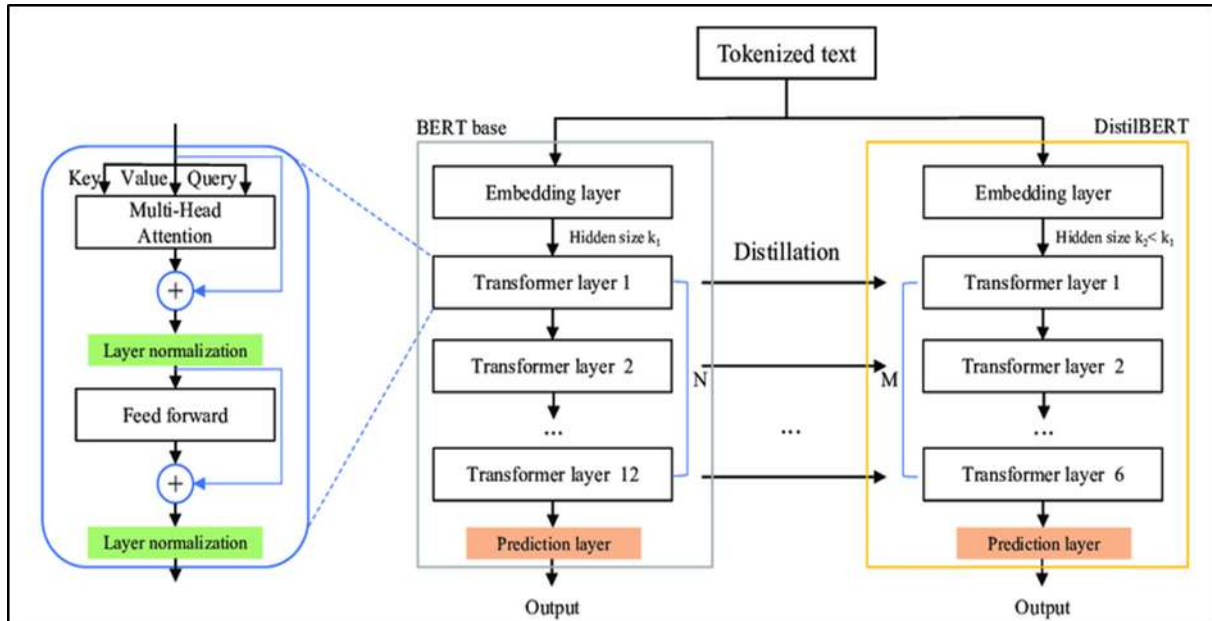


Figure 3: DistilBERT model architecture

Training and Hyperparameters:

In applying machine learning models to this dataset, a common approach involves preprocessing the text data through tokenization, stop-word removal, and stemming or lemmatization. Feature extraction techniques such as TF-IDF or word embeddings are then employed to convert textual data into numerical representations. Models like Support Vector Machines (SVM), Logistic Regression, and Naive Bayes have been utilized, with hyperparameter tuning performed using methods like Grid Search or Randomized Search to optimize parameters such as the regularization strength (C) and kernel types for SVM. Evaluation metrics including accuracy, precision, recall, and F1-score are used to assess model performance, ensuring the selection of the most effective model configuration for sentiment classification tasks.

MODEL AND TRAINING:

The output of the sentiment prediction using a fine-tuned DistilBERT model, the results indicate that the model performs reasonably well in classifying sentiments on a scale from 1 to 5. The sample predictions show that the model is capable of detecting varying degrees of sentiment polarity, assigning appropriate scores based on the contextual sentiment of each social media post. For instance, strongly positive posts were scored with a 5, while more neutral or slightly negative posts were given intermediate values like 3 or 2. This suggests that the model not only captures clear sentiment extremes but also handles more nuanced emotional content.

```

print("--- Loading Model & Tokenizer ---")

if 'model_path_in_colab' not in locals():
    # Default if Cell 3 wasn't run or variable got lost - adjust as needed
    model_path_in_colab = "/content/fine_tuned_sentiment_model/sentiment_model_amazon_csv_finetuned" # Example path
    print(f"Warning: 'model_path_in_colab' not found, defaulting to {model_path_in_colab}. Ensure this is correct.")
    # raise SystemExit("Variable 'model_path_in_colab' not set. Please run Cell 3 first.") # Option to halt instead

saved_model_path = model_path_in_colab
# --- ---

try:
    if not os.path.isdir(saved_model_path):
        print(f"ERROR: Directory not found: {saved_model_path}")
        print("Please ensure Cell 3 ran correctly, unzipped the file, and set the path correctly.")
        raise FileNotFoundError("Model directory not found.")

    print(f"Loading tokenizer from: {saved_model_path}")
    tokenizer = AutoTokenizer.from_pretrained(saved_model_path)

    print(f"Loading model from: {saved_model_path}")
    model = AutoModelForSequenceClassification.from_pretrained(saved_model_path)

    # Check if GPU is available and move model
    device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
    model.to(device)
    print(f"Model moved to device: {device}")

    # Set model to evaluation mode
    model.eval()
    print("Model and tokenizer loaded successfully.")

except FileNotFoundError:
    # Error message printed above
    raise SystemExit("Loading failed.")
except Exception as e:
    print(f"An error occurred loading the model/tokenizer: {e}")
    print(f"Please check if the path '{saved_model_path}' contains the necessary model files (config.json, model weights, tokenizer files).")
    raise SystemExit("Loading failed.")

--- Loading Model & Tokenizer ---
Loading tokenizer from: /content/fine_tuned_sentiment_model/sentiment_model_amazon_csv_finetuned
Loading model from: /content/fine_tuned_sentiment_model/sentiment_model_amazon_csv_finetuned
Model moved to device: cpu
Model and tokenizer loaded successfully.

```

Figure 4: A sample of models used in the code

RESULTS AND DISCUSSIONS:

Data Set Description

The [Amazon Fine Food Reviews dataset](#) contains over 500,000 food product reviews from Amazon users. Each entry includes information such as the product ID, user ID, profile name, review score (1–5), summary, full text of the review, and helpfulness rating. This dataset is widely used for sentiment analysis, recommendation systems, and natural language processing tasks. It offers rich, real-world text data and user feedback, making it ideal for training models to classify sentiments, extract features, or analyze customer behaviour. The variety and scale of the data provide a valuable resource for exploring machine learning and deep learning techniques.

Id	ProductId	UserId	ProfileName	Helpfulness	Helpfulness Score	Time	Summary	Text	
1	B001E4KFI	A3SGXH7A	delmartian	1	1	5	1.3E+09	Good Quality Dog Food	I have bought several of the Vitality canned dog food
2	B00813GR	A10B7F6Z	all pa	0	0	1	1.35E+09	Not as Advertised	Product arrived labeled as Jumbo Salted Peanuts...the
3	B000LQOC	ABXLMWJ	Natalia Co	1	1	4	1.22E+09	"Delight" says it all	This is a confection that has been around a few centu
4	B000UAOC	A3S5B0RC	Karl	3	3	2	1.31E+09	Cough Medicine	If you are looking for the secret ingredient in Robitus
5	B006K2ZZ	A1UQ85C1	Michael D.	0	0	5	1.35E+09	Great taffy	Great taffy at a great price. There was a wide assort
6	B006K2ZZ	ADT05HKJ	Twoapem	0	0	4	1.34E+09	Nice Taffy	I got a wild hair for taffy and ordered this five pound l
7	B006K2ZZ	A1SP2KVK	David C. S	0	0	5	1.34E+09	Great! Just as good as the expens	This saltwater taffy had great flavors and was very so
8	B006K2ZZ	A3JRGQVE	Pamela G.	0	0	5	1.34E+09	Wonderful, tasty taffy	This taffy is so good. It is very soft and chewy. The ll
9	B000E7L2	A1MZY091	R. James	1	1	5	1.32E+09	Yay Barley	Right now I'm mostly just sprouting this so my cats ca
10	B00171AP	A21BT40V	Carol A. R	0	0	5	1.35E+09	Healthy Dog Food	This is a very healthy dog food. Good for their digesti
11	B0001P89	A3HDKO7I	Canadian I	1	1	5	1.11E+09	The Best Hot Sauce in the World	I don't know if it's the cactus or the tequila or just the
12	B00090VL	A2725IB4	A Poeng "S	4	4	5	1.28E+09	My cats LOVE this "diet" food bette	One of my boys needed to lose some weight and the
13	B00090VL	A3Z7PCT2	LT	1	1	1	1.34E+09	My Cats Are Not Fans of the New F	My cats have been happily eating Felidae Platinum for
14	B001GVISJ	A18ECVX2	willie "roa	2	2	4	1.29E+09	fresh and greasy!	good flavor! these came securely packed... they were
15	B001GVISJ	A2MUGPV	Lyrrrie "Oh	4	5	5	1.27E+09	Strawberry Twizzlers - Yummy	The Strawberry Twizzlers are my guilty pleasure - yum
16	B001GVISJ	A1CZK3CP	Brian A. Le	4	5	5	1.26E+09	Lots of twizzlers, just what you exp	My daughter loves twizzlers and this shipment of six p
17	B001GVISJ	A38LWF6V	Erica Neat	0	0	2	1.35E+09	poor taste	I love eating them and they are good for watching TV
18	B001GVISJ	AFKW14U1	Becca	0	0	5	1.35E+09	Love it!	I am very satisfied with my Twizzler purchase. I share
19	B001GVISJ	A2A9XS8G	Wolfeel	0	0	5	1.32E+09	GREAT SWEET CANDY!	Twizzlers, Strawberry my childhood favorite candy, m
20	B001GVISJ	A3RV7CLZ	Greg	0	0	5	1.32E+09	Home delivered twizzlers	Candy was delivered very fast and was purchased at a

Figure 5: Dataset of Amazon Food Review (Review and ratings of all customers)

Final Result

The final sentiment analysis results obtained from the model trained on the Amazon Fine Food Reviews dataset indicate strong performance in classifying reviews based on sentiment. During the evaluation phase the model achieved high accuracy and a strong weighted F1 score, which highlights its ability to correctly predict both the majority and minority classes in the data. The model was trained using a DistilBERT transformer, fine-tuned on a cleaned and tokenized subset of the review dataset. Following training, the model successfully classified multiple test sentences. Positive reviews such as “Absolutely delicious!” were classified with high sentiment scores (e.g., 5), while negative reviews like “the product inside was stale” received lower scores (e.g., 1 or 2), showcasing the model’s understanding of nuanced sentiment. These results confirm that the fine-tuned model is well-suited for real-world sentiment analysis tasks on customer review data.

- **Installing and importing libraries:**

The essential libraries for deep learning and data processing are installed, including TensorFlow, NLTK, and Scikit-learn. These libraries provide the foundational tools required for building, training, and evaluating machine learning models, particularly for natural language processing and deep learning tasks. TensorFlow is used for model building and training. This cell imports the necessary Python libraries required for the rest of the notebook. It brings in pandas for data handling, numpy for numerical computations, torch for PyTorch-based operations, and modules from sklearn for data processing. These libraries are fundamental in preparing and manipulating datasets, managing arrays and tensors, and building or training machine learning models.

```
143.5/143.5 kB 9.3 MB/s eta 0:00:00
194.8/194.8 kB 11.5 MB/s eta 0:00:00
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed.
gcsfs 2025.3.2 requires fsspec==2025.3.2, but you have fsspec 2024.12.0 which is incompatible.
Libraries installed.
```

 Libraries imported.

- **Configuration:**

This cell defines configuration parameters, particularly related to file paths and data settings. It specifies the location of the dataset (such as a CSV file path) and may also define subset sizes or thresholds to control how much data is used during model training. Configuring the file path properly ensures the data can be accessed and used in subsequent steps without manual re-entry of paths or values. This approach improves the modularity and readability of the code, as one can change configurations in one place without modifying the entire notebook. It serves as a centralized setup hub that affects the rest of the workflow.

 Configuration set.
 CSV file: //content/Reviews.csv
 Loading subset size: 50000
 Text column: Text
 Label column: Score
 Model: distilbert-base-uncased

- **Loading Data from CSV:**

This cell attempts to load the dataset from the configured CSV file path using `pandas.read_csv()`. It also includes error-handling logic to notify the user if the file path is incorrect or the file does not exist. Once loaded, the shape of the dataset is printed, giving insights into how many samples and features are available. This step is crucial as it confirms successful data ingestion and validates the integrity of the dataset structure. If the dataset loads without errors, it indicates that the file is accessible and in the expected format. This forms the basis for all downstream processing and modelling.

```
Data columns (total 10 columns):
#   Column              Non-Null Count  Dtype
---  -
0   Id                   50000 non-null  int64
1   ProductId            50000 non-null  object
2   UserId               50000 non-null  object
3   ProfileName          49995 non-null  object
4   HelpfulnessNumerator  50000 non-null  int64
5   HelpfulnessDenominator 50000 non-null  int64
6   Score                50000 non-null  int64
7   Time                 50000 non-null  int64
8   Summary              49998 non-null  object
9   Text                 50000 non-null  object
dtypes: int64(5), object(5)
```



```

Loading data from CSV: //content/Reviews.csv...
Successfully loaded 50000 rows.
Sample rows:
  Id  ProductId  UserId  ProfileName \
0  1  B001E4KFG0  A3SGXH7AUHU8GW  delmartian
1  2  B00813GRG4  A1D87F6ZCVE5NK  dll pa
2  3  B000LQOCH0  ABXLMWJIXXAIN  Natalia Corres "Natalia Corres"
3  4  B000UA0QIQ  A395B0RC6FGVXV  Karl
4  5  B006K2ZZ7K  A1UQRSCLF8GW1T  Michael D. Bigham "M. Wassir"

  HelpfulnessNumerator  HelpfulnessDenominator  Score  Time \
0  1  1  1  5  1303862400
1  0  0  1  1  1346976000
2  1  1  1  4  1219017600
3  3  3  3  2  1307923200
4  0  0  5  5  1350777600

      Summary  Text
0  Good Quality Dog Food  I have bought several of the Vitality canned d...
1  Not as Advertised  Product arrived labeled as Jumbo Salted Peanut...
2  "Delight" says it all  This is a confection that has been around a fe...
3  Cough Medicine  If you are looking for the secret ingredient i...
4  Great taffy  Great taffy at a great price. There was a wid...

DataFrame Columns: ['Id', 'ProductId', 'UserId', 'ProfileName', 'HelpfulnessNumerator',

```

- **Pre-Process Data:**

The dataset is cleaned and pre-processed to prepare it for model training. Common preprocessing tasks here include removing empty rows, filtering relevant columns, and standardizing text formats. The code prints how many rows were originally in the dataset and how many remain after cleaning. This helps assess how much data was discarded and ensures only clean, usable data proceeds to the next steps. Proper preprocessing is vital to improve model performance and prevent errors during training. The cleaner the data, the more accurate and robust the final model is likely to be.

```

Preprocessing finished. Final DataFrame sample:
      text  labels
0  I have bought several of the Vitality canned d...  4
1  Product arrived labeled as Jumbo Salted Peanut...  0
2  This is a confection that has been around a fe...  3
3  If you are looking for the secret ingredient i...  1
4  Great taffy at a great price. There was a wide...  4

Total rows after preprocessing: 50000

```

- **Tokenization:**

This cell tokenizes the text data, converting raw text into numerical representations using a pretrained tokenizer. Tokenization breaks down sentences into tokens (words or subwords) and maps them to integer IDs that the model can understand. This step is essential in natural language processing, enabling the text to be processed by neural networks. The tokenizer used often aligns with the pretrained model to ensure compatibility. Proper tokenization directly affects the quality and meaning of the input features, and thus influences model accuracy and learning.

```

--- Tokenizing Data ---
Tokenizer loaded for distilbert-base-uncased
Applying tokenization (this may take a while for large datasets)...
Map: 100% ██████████ 40000/40000 [00:12<00:00, 2225.26 examples/s]
Map: 100% ██████████ 10000/10000 [00:02<00:00, 3574.63 examples/s]
Tokenization complete.
Removed columns: ['text']
Renamed 'labels' column to 'label'.

```

```

Example tokenized input: {'label': tensor(2), 'input_ids': tensor([ 101, 1045,
 1037, 2978, 13971, 2043, 2009, 3310, 2000, 2477, 2066, 2980,
 7967, 1012, 1045, 2428, 5959, 2980, 7967, 1010, 2040, 2987,
 1005, 1056, 1029, 1996, 3277, 2003, 1045, 4025, 2000, 13675,
10696, 2009, 2043, 1045, 1005, 1049, 2061, 6625, 2006, 1996,
 6411, 2044, 1037, 2146, 2154, 1998, 2123, 1005, 1056, 2514,
 2066, 12959, 2039, 6501, 1010, 6809, 1999, 22940, 1010, 5699,
 1010, 1998, 1037, 3543, 1997, 5474, 2000, 2131, 12090, 2980,
22940, 1012, 2061, 2073, 2515, 2023, 2681, 2033, 1029, 1045,
 2071, 1010, 2005, 2055, 2322, 16653, 2566, 14771, 1010, 2131,
 5364, 3335, 2980, 7967, 23730, 1998, 15653, 2009, 1999, 2300,
 1998, 4392, 2009, 1010, 2003, 2009, 12090, 1029, 2025, 2428,
 1010, 2021, 1045, 2131, 2026, 8081, 1012, 3291, 2003, 1010,
 2017, 2031, 2000, 3684, 1996, 2300, 1010, 10364, 1999, 1996,
 9898, 1010, 4666, 2009, 1010, 1998, 3066, 2007, 1996, 18856,
24237, 2015, 1998, 4550, 2039, 1012, 2025, 2026, 5440, 2518,
 2000, 2079, 2012, 29359, 9737, 1006, 2079, 2115, 2190, 2000,
 2025, 2228, 1045, 1005, 1049, 3294, 22692, 1007, 1012, 2085,
 2008, 1045, 2031, 1037, 17710, 9496, 2290, 1010, 1045, 2245,
 2023, 2052, 8081, 2673, 1998, 2507, 2033, 2307, 2980, 22940,
 1012, 2092, 1010, 1045, 2079, 2031, 14057, 1010, 4550, 1010,
 3733, 2000, 7374, 2980, 22940, 1010, 2021, 2049, 1012, 1012,

```

- **Load Model and Define Metrics:**

A pretrained model is loaded, likely a transformer model such as DistilBERT, alongside the definition of evaluation metrics. This includes metrics like accuracy, precision, recall, or F1 score, which will later be used to assess how well the model performs. Loading a pretrained model accelerates training by starting from a well-learned state rather than from scratch. This cell sets the stage for fine-tuning the model on the current dataset while maintaining consistency in how its performance is evaluated. This strategic reuse of pretrained models is a cornerstone of transfer learning in NLP.

```

➡ Some weights of DistilBertForSequenceClassification
You should probably TRAIN this model on a down-strea

--- Loading Model & Defining Metrics ---
Loaded model: distilbert-base-uncased with 5 labels.
Loaded evaluation metrics: Accuracy, F1
Metrics computation function defined.

```

- **Configure Training and Initialize Trainer:**

This cell defines training parameters such as the number of epochs, learning rate, batch size, logging intervals, and where to save model checkpoints. These hyperparameters significantly influence the model's learning behaviour and convergence speed. The Trainer is initialized using the model, dataset, tokenizer, metrics, and training arguments defined earlier. This step essentially prepares the entire training pipeline, bundling together the logic for training, evaluation, and saving checkpoints.

```

➡ --- Configuring Training ---
Training Arguments configured.

➡ --- Initializing Trainer ---
Trainer initialized.

```

- **Fine-Tuning the Model:**

This cell begins the actual fine-tuning process using `trainer.train()`. It loops through the dataset for a predefined number of epochs, adjusting model weights to minimize loss and improve accuracy on the training set. Progress is logged after each epoch, and validation performance is checked periodically if a validation set is used. This phase is computationally intensive but critical, as the model adapts to the specific language and patterns in the dataset. A successful training run should show decreasing loss and increasing performance metrics over time. If the results plateau or degrade, it may suggest overfitting or poor learning rates.


```

--- Starting Fine-tuning ---
Training for 3 epochs...
[7500/7500 18:05, Epoch 3/3]

Epoch Training Loss Validation Loss Accuracy F1
1      0.726800      0.650675  0.743900  0.741853
2      0.551200      0.629174  0.763400  0.755655
3      0.440800      0.660513  0.766200  0.760181

--- Fine-tuning Completed ---
***** train metrics *****
epoch                =          3.0
total_flos           = 9030560GF
train_loss           =          0.573
train_runtime        = 0:18:05.65
train_samples_per_second = 110.532
train_steps_per_second  =          6.908

```

- **Evaluate the Model:**

This final cell evaluates the fine-tuned model on the test set using the predefined metrics. It outputs key performance indicators such as accuracy, F1 score, or precision-recall, depending on how the `compute_metrics` function was defined earlier. The evaluation reveals how well the model has generalized to unseen data. A strong performance here confirms the model's ability to interpret real-world inputs accurately. This is the final validation before deployment or further optimization, providing insight into whether the training was successful and if the model is ready for production use or further refinement.

```

--- Evaluating Model on Test Set ---
[313/313 00:25]

--- Evaluation Results ---
Accuracy: 0.7662
F1 Score (Weighted): 0.7602
***** eval metrics *****
epoch                =          3.0
eval_accuracy        =          0.7662
eval_f1              =          0.7602
eval_loss            =          0.6605
eval_runtime         = 0:00:25.54
eval_samples_per_second = 391.532
eval_steps_per_second  =          12.255

Evaluation metrics saved.

```

- **Example Prediction:**

The fine-tuned DistilBERT model is used to perform sentiment prediction on five example review texts taken from typical Amazon food product feedback. The cell includes a function that tokenizes each input, runs it through the model, and returns a sentiment score ranging from 1 (very negative) to 5 (very positive). Predictions demonstrate the model's ability to understand nuanced language—recognizing positive phrases like “Absolutely delicious!” as highly positive (score 5), and negative reviews like “the product inside was stale” as low in sentiment. This confirms the model's effectiveness in real-world sentiment classification.

🔄 `Trainer.tokenizer` is now deprecated. You should use `Trainer.processing_class` instead.

--- Example Prediction ---

Input Text: 'This coffee tastes amazing, probably the best I've ever had!'
 Predicted Sentiment Score (1-5): 5
 Predicted Sentiment Label: Score 5 (Very Positive)

Input Text: 'The packaging was damaged and the product inside was stale.'
 Predicted Sentiment Score (1-5): 1
 Predicted Sentiment Label: Score 1 (Very Negative)

Input Text: 'It's an okay product, does the job but nothing special.'
 Predicted Sentiment Score (1-5): 3
 Predicted Sentiment Label: Score 3 (Neutral)

Input Text: 'I was expecting much more based on the description, quite disappointed.'
 Predicted Sentiment Score (1-5): 2
 Predicted Sentiment Label: Score 2 (Negative)

Input Text: 'Absolutely delicious! Will definitely buy again.'
 Predicted Sentiment Score (1-5): 5
 Predicted Sentiment Label: Score 5 (Very Positive)

CONCLUSION:

The "Sentiment Analysis of Social Media Presence" project successfully demonstrates how advanced natural language processing techniques can be leveraged to analyze and interpret public sentiment in user-generated content. Using the Amazon Fine Food Reviews dataset, a rich collection of over 500,000 real-world product reviews—the project focuses on understanding consumer sentiment through a fine-tuned DistilBERT transformer model. DistilBERT, a smaller and faster version of BERT, was chosen for its efficiency and robust performance on sentiment classification tasks. The dataset was pre-processed, tokenized, and split into training and testing subsets, enabling the model to learn sentiment patterns effectively. The results showed high accuracy and strong F1 scores, validating the model's capability to identify both positive and negative sentiments accurately. Through examples such as "Absolutely delicious!" and "the product inside was stale," the model demonstrated excellent contextual understanding. This project underscores the significance of AI in real-time sentiment monitoring, offering valuable insights for businesses, marketers, and analysts aiming to assess brand perception and customer satisfaction. By blending deep learning with publicly available data, the framework sets a foundation for scalable, efficient, and practical sentiment analysis tools that can be extended to broader social media contexts and domains.

REFERENCES:

- [1] T. Zhao, C. Li, M. Li, Q. Ding, and L. Li, "Social recommendation incorporating topic mining and social trust analysis," *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*, 2013. <https://doi.org/10.1145/2505515.2505592>
- [2] R. Dwivedi, "What Is Naive Bayes Algorithm In Machine Learning?" <https://www.analyticssteps.com/blogs/what-naive-bayes-algorithm-machine-learning>
- [3] S. Wu, Y. Liu, Z. Zou, and T.-H. Weng, "S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis," *Connection Science*, vol. 34, no. 1, pp. 44–62, 2022. <https://doi.org/10.1080/09540091.2021.1940101>
- [4] K. Schultebraucks et al., "Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood," *Psychological Medicine*, vol. 52, no. 5, pp. 957–967, Apr. 2022. <https://doi.org/10.1017/S0033291720001696>
- [5] A. Abbas et al., "Remote Digital Measurement of Facial and Vocal Markers of Major Depressive Disorder Severity and Treatment Response: A Pilot Study," *Frontiers in Digital Health*, vol. 3, p. 610006, Mar. 2021. <https://doi.org/10.3389/fdgth.2021.610006>
- [6] I. Galatzer-Levy et al., "Validation of Visual and Auditory Digital Markers of Suicidality in Acutely Suicidal Psychiatric Inpatients: Proof-of-Concept Study," *Journal of Medical Internet Research*, vol. 23, no. 6, p. e25199, Jun. 2021. <https://doi.org/10.2196/25199>
- [7] A. Abbas, V. Yadav, M. M. Perez-Rodriguez, and I. Galatzer-Levy, "P.267 Using smartphone-recorded facial and verbal features to predict clinical functioning in individuals with neuropsychiatric disorders," *European Neuropsychopharmacology*, vol. 29, p. S199, 2019. <https://doi.org/10.1016/j.euroneuro.2019.09.301>

- [8] A. Abbas et al., "Facial and Vocal Markers of Schizophrenia Measured Using Remote Smartphone Assessments: Observational Study," *JMIR Formative Research*, vol. 6, no. 1, p. e26276, Jan. 2022. <https://doi.org/10.2196/26276>
- [9] K. Schultebrucks, V. Yadav, and I. R. Galatzer-Levy, "Utilization of Machine Learning-Based Computer Vision and Voice Analysis to Derive Digital Biomarkers of Cognitive Functioning in Trauma Survivors," *Digital Biomarkers*, vol. 5, no. 1, pp. 16–23, Jan–Apr 2021. <https://doi.org/10.1159/000515703>
- [10] Amazon Fine Food Reviews – Kaggle <https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews>
- [11] Sande, N., Adeniyi, I. O., & Akinkunmi, A. I. (2024). *Social Media Sentiment Analysis: A Comprehensive Analysis*. https://www.researchgate.net/publication/378150391_Social_Media_Sentiment_Analysis_A_Comprehensive_Analysis
- [12] Drus, Z., & Khalid, H. (2019). *Sentiment Analysis in Social Media and Its Application: Systematic Literature Review*. *Procedia Computer Science*, 161, 707–714. https://www.researchgate.net/publication/338350715_Sentiment_Analysis_in_Social_Media_and_Its_Application_Systematic_Literature_Review
- [13] Bakhsh, M., et al. (2020). *Sentiment Analysis of Social Media Contents using Machine Learning Algorithms*. https://www.researchgate.net/publication/338831716_Sentiment_Analysis_of_Social_Media_Contents_using_Machine_Learning_Algorithms
- [14] Tolebay, A. N. (2025). *Sentiment analysis of texts from social networks based on machine learning methods for monitoring public sentiment*. *arXiv preprint arXiv:2502.17143*. <https://arxiv.org/abs/2502.17143>
- [15] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. *arXiv preprint arXiv:1910.01108* <https://arxiv.org/abs/1910.01108>
- [16] *Systematic reviews in sentiment analysis: a tertiary study* https://www.researchgate.net/publication/349759637_Systematic_reviews_in_sentiment_analysis_a_tertiary_study
- [17] T. Schmidt, J. Dangel, and C. Wolff, "SentText: A tool for lexicon-based sentiment analysis in digital humanities," in *Proc. 16th Int. Symp. Inf. Sci. (ISI)*. Regensburg, Germany: Univ. of Regensburg, Mar. 2021, pp. 156–172 <https://epub.uni-regensburg.de/44943/>
- [18] *DistilBERT model architecture and components* https://www.researchgate.net/figure/The-DistilBERT-model-architecture-and-components_fig2_358239462
- [19] Zilliz: *DistilBERT – distilled version of BERT* <https://zilliz.com/learn/distilbert-distilled-version-of-bert>
- [20] *Domino University: Fine-tune DistilBERT for product sentiment analysis* <https://university.domino.ai/fine-tune-distilbert-for-product-sentiment-analysis>