Sentiment Analysis Project Report

DATA 5100 - Group 3 Project

Edwin Okwor, Anushka Naidu and Jimmy Nam

December 11, 2025

1. Abstract

This project explores the ability of a fine-tuned Transformer model to accurately classify the sentiment of social media posts into three categories: Positive, Negative, and Neutral. It also examines how sentiment relates to engagement metrics, specifically likes and retweets. Utilizing a dataset of 717 labeled posts, a DistilBERT-based uncased model is fine-tuned and evaluated for sentiment classification performance.

2. Introduction

Have you ever wondered, while scrolling through a platform like Twitter—now known as X—why some posts receive significantly more likes and retweets than others? Could it be the sentiment of the post, or might other factors be at play—such as who is posting it, whether the content is humorous, or whether the post is designed to elicit engagement?

Sentiment analysis has become increasingly vital in the field of data science, as businesses, researchers, and individuals seek to understand how people perceive products, services, brands, and social content. With millions of users sharing opinions, experiences, and reactions online every day, sentiment analysis provides valuable insights into public attitudes and emerging trends. By detecting shifts in satisfaction or engagement, organizations and content creators can identify which posts or offerings resonate most with their audience. This information helps guide strategic decisions, optimize interactions, and improve user experiences across a variety of domains.

This project focuses on answering two key questions:

1. Can a fine-tuned Transformer model reliably classify social media text into Positive, Neutral, and Negative sentiments?
2. Is there a statistically significant relationship between sentiment and engagement metrics (likes and retweets)?

To address these questions, we implement a complete analytical pipeline that includes data cleaning, grouping, ambiguity reduction, tokenization, fine-tuning a Transformer-based model, and evaluating engagement patterns. For effective text analysis, we selected DistilBERT due to its strong contextual understanding, reduced computational cost, and consistently strong performance across a wide range of Natural Language Processing (NLP) tasks.

3. Theoretical Background

3.1 Sentiment Analysis

Sentiment analysis is a key area within Natural Language Processing that focuses on identifying the emotional tone present in written text. Early approaches relied on methods such as lexico-based scoring and bag-of-words representations, both of which treated text as a collection of discrete terms. These techniques overlooked the influence of word order, contextual meaning, and subtle emotional cues. As a result, they often struggled with social media content, which typically contains informal language, sarcasm, abbreviations, emojis, and short expressive statements. The limitations of these traditional models became more apparent as researchers began relying on sentiment analysis to understand public attitudes, customer experiences, and patterns of online engagement. Modern approaches, therefore, shifted toward machine learning and deep learning frameworks that can recognize patterns in emotional expression beyond simple word counts.

3.2. Context in Language

A major development in Natural Language Processing came with the emergence of contextual language models, especially those based on the Transformer architecture. Instead of processing text sequentially, Transformer models use self-attention to examine relationships among all words in a sentence at the same time. This allows the model to capture long-range dependencies and shifts in meaning that depend on context. For example, a phrase that appears positive in isolation may take on a negative tone when paired with a different surrounding phrase. Because social media posts often rely on implied meaning, indirect expression, or conversational nuance, the ability to interpret context is essential for accurate sentiment classification.

3.3. Transformer Models

Pretrained Transformer models such as DistilBERT have greatly improved performance on sentiment analysis tasks because they are first trained on large general-purpose text corpora. During this initial training, the models learn broad linguistic patterns that include grammar, semantics, and typical patterns of emotional expression. Fine-tuning allows researchers to adapt these pretrained models to the specific characteristics of a new dataset. In the context of sentiment analysis on social media, fine-tuning allows researchers to adapt these pretrained models to the specific characteristics of a new dataset. In the context of sentiment analysis on social media, fine-tuning helps the model learn how sentiment is conveyed within very short

posts that may combine text with emojis or informal phrasing. DistilBERT in particular offers strong performance with reduced computational requirements, making it suitable for datasets of modest size.

3.4. Engagement Theory

Engagement represents reactions such as likes and retweets that reflect a user's response to a post. Prior research suggests that emotional content often attracts more interaction because people respond more readily to expressive or positive language. Even so, engagement is influenced by many factors, which means sentiment alone cannot explain all variation in user behavior.

3.5. Statistical Modeling

Ordinary Least Squares regression provides a simple and interpretable method for testing how sentiment relates to engagement. By comparing positive and negative posts to a neutral baseline, the model can estimate whether emotional tone leads to higher or lower interaction. This framework connects the sentiment predictions produced by the Transformer model to measurable differences in user behavior.

4. Methodology

The methodology used includes three(3) major components: preparing the dataset, building a sentiment classification model, and running a statistical analysis to understand how sentiment affects engagement.

4.1 Dataset Overview

The dataset was obtained from Kaggle and consists of 732 initial social media posts, with each containing:

- Post Text Content
- A sentiment label
- Engagement metrics (likes, retweets)
- Metadata(timestamp, user, platform, country)

The dataset can be accessed here: https://www.kaggle.com/datasets/kashishparmar02/social-media-sentiments-analysis-dataset.

4.2. Data Cleaning and Preparation

The dataset included the text of each post, its assigned sentiment label, and corresponding engagement metrics. During preprocessing, the data was cleaned by removing missing or NaN

values, and all sentiment labels were standardized to lowercase and stripped of excess whitespace.

To simplify analysis, all emotion labels were consolidated into three primary categories: positive, neutral, and negative. Any ambiguous labels that did not fit clearly into these groups were removed.

4.3. Training/Validation/Test Split

The finalized sentiment classes were then encoded numerically, and the dataset was split into training, validation, and test sets using stratified sampling to preserve the original class distribution.

4.4 Tokenization

Transformer models such as DistilBERT cannot read raw text directly. It requires text to be converted into numerical tokens that represent: words, punctuations, and special model-specific markers. To achieve this, we used the HuggingFace DistilBERT tokenizer - a text-processing component designed specifically for the DistilBERT model-to transform our input data into the appropriate tokenized format that the model can interpret.

4.5 Class Weighting

After grouping the datasets by sentiment category, we observed a significant class imbalance. The positive class contained 472 samples, the negative class had 177, and the neutral class had only 68. To address this disproportionate distribution, we applied class weights, which increase the loss contribution of minority classes during training. Because neutral posts made up the smallest portion of the dataset, they received the weight. This ensured that the model paid more attention to the neutral class and learned to classify all sentiments more fairly.

In order to classify the sentiments more fairly, the class weights penalize the model more heavily when it gets the minority class wrong. Because the minority classes-neutral have fewer samples, the model would normally learn less about them. Class Weighting fixes this by increasing the loss for those classes. So if the model misclassifies a neutral post(the smallest class), the error contributes more to the total loss than misclassifying a positive one.

This often forces the model to take the minority classes seriously and prevents from predicting the majority class every time.

4.6 Model Training

The model was trained for 5 epochs with a batch size of 16, a learning rate of 2e-5, and a warmup ratio of 0.1 to stabilize the early stages of finetuning. Training was conducted using a class-weighted cross-entropy loss function that was implemented through a custom WeightedTrainer.

To prevent overfitting, early stoppage was incorporated using a patience of two(2), which monitored the validation loss and terminated training automatically when the model was no longer improving. We also enabled an evaluation strategy during training so that the performance metrics were computed at the end of each epoch using the validation dataset. This was critical to ensure that the trainer tracked improvements and automatically saved the best-performing model based on the lowest validation loss.

Altogether, the training configuration ensured stable optimization, mitigated class imbalance, and facilitated the selection of an optimal model.

4.7 Model Evaluation

The trained model was evaluated using accuracy, precision, recall, and F1-scores. A confusion matrix was also used to show how well the model separated the three sentiment categories and where most errors occurred, especially within the neutral group. The final predicted labels were then used as the independent variable in the engagement analysis.

4.8 Engagement Variable Construction

For simplicity and to capture overall user interaction, a new variable—**engagement**—was created by combining the number of likes and retweets for each post. This aggregated metric serves as the dependent variable in the regression model, providing a single measure of total engagement.

4.9. Statistical Modeling: OLS Regression

An Ordinary Least Squares regression model was used to test whether sentiment predicts engagement. Dummy variables were created for positive and negative sentiment, and neutral was used as the reference category. The model evaluated whether engagement for these categories differed from neutral posts. R^2, the F-statistic, and the coefficients were used to interpret the strength and significance of the relationship between sentiment and engagement.

4.10. Exploratory Data Analysis

Before running the regression, visualizations were created to understand the distribution of sentiment predictions, the average engagement across categories, and how engagement is distributed within each group. These plots helped guide the interpretation of the regression results.

The exploratory data analysis focused on understanding the distribution of predicted sentiment categories and examining engagement patterns across negative, neutral, and positive posts. The goal of the EDA was to identify initial trends in the data and build intuition before conducting statistical modeling.

5. Computational Results

5.1 Sentiment Distribution After Cleaning

Bar Chart displaying Counts

A bar chart was generated to visualize the distribution of sentiment categories after the data-cleaning process. This visualization provided a clear overview of how the posts were distributed across the three(3) sentiment classes. The chart revealed a noticeable class imbalance: the dataset contained substantially more Positive and Negative posts, while the Neutral category was significantly underrepresented. This imbalance later informed the decision to apply class-weighted loss during model training to ensure that all sentiment classes were learned effectively.

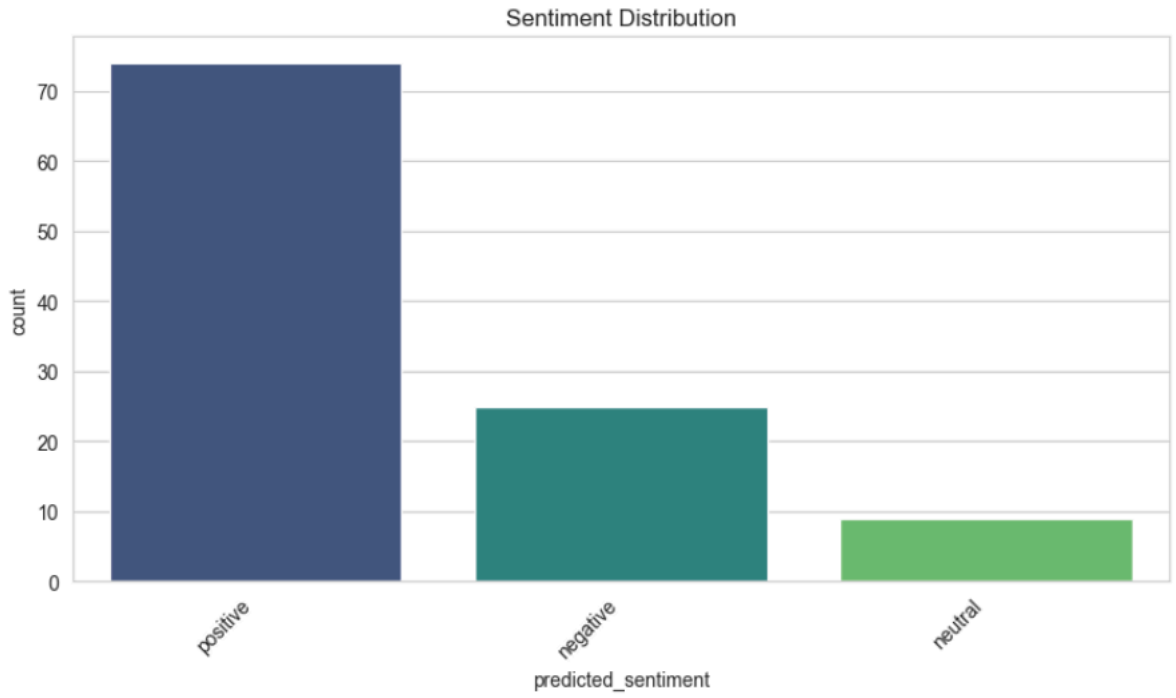


Figure 1. Sentiment Distribution

Figure 1 shows that the majority of posts were classified as **positive**, followed by **negative**, with **neutral** being the least frequent category. This suggests that the text content analyzed in the dataset tends to lean toward positive expression. The comparatively smaller neutral group also indicates that fewer posts fall into an emotionally neutral or mixed category, which is common in social media datasets where people often express clear emotional tones.

Engagement Summary Table

To better understand how engagement differs across sentiment categories, the dataset was grouped by predicted sentiment, and summary statistics were calculated for likes and retweets. The resulting table includes the total and average number of likes and retweets for negative, neutral, and positive posts.

	predicted_sentiment	likes_Total	likes_Average	retweets_Total	retweets_Average
0	negative	952.0	38.080000	477.0	19.080000
1	neutral	385.0	42.777778	194.0	21.555556
2	positive	3447.0	46.581081	1726.0	23.324324

Table 1. Engagement Summary by Sentiment

Table 1 shows clear differences in user interaction depending on the sentiment of the post.

- **Positive posts** have the highest totals and averages for both likes and retweets.
- **Neutral posts** receive moderate engagement, slightly lower than positive posts but still higher than negative posts on average.
- **Negative posts** show the lowest total and average engagement across categories.

This table provides a useful numerical foundation that helps highlight engagement patterns before viewing them visually.

Average Likes and Retweets per Sentiment Category

A grouped bar chart was created using the average likes and average retweets from the engagement summary table. In this chart, the x-axis represents the sentiment categories, while the y-axis represents the average number of interactions. Two bars are shown for each category, one for likes and one for retweets, allowing visual comparison across both engagement metrics.

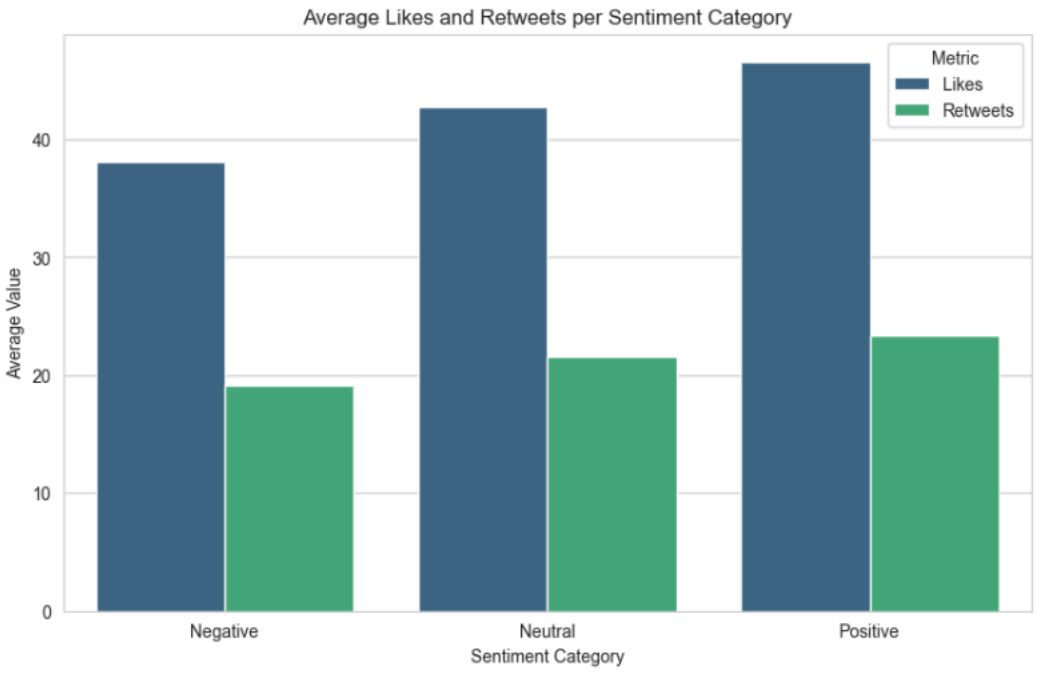


Figure 2. Average Likes and Retweets per Sentiment Category

Figure 2 shows a clear pattern in engagement behavior. Posts with **positive sentiment** receive the highest average likes and retweets, indicating that users interact more frequently with

uplifting or emotionally positive content.
Neutral posts show strong engagement levels as well, with averages close to those of positive posts. In contrast, **negative posts** receive noticeably lower engagement, suggesting that users are less likely to interact with content expressing negative emotions.

Overall, the visualization highlights a consistent trend in which more positive emotional expression corresponds to increased engagement, while negative sentiment appears to reduce or limit user interaction.

Engagement Density by Sentiment Category

To examine the distribution of engagement across sentiment groups, a kernel density curve was created using total engagement as the continuous variable. The density plot allows comparison of the distribution shapes and highlights where engagement values tend to cluster for each sentiment category.

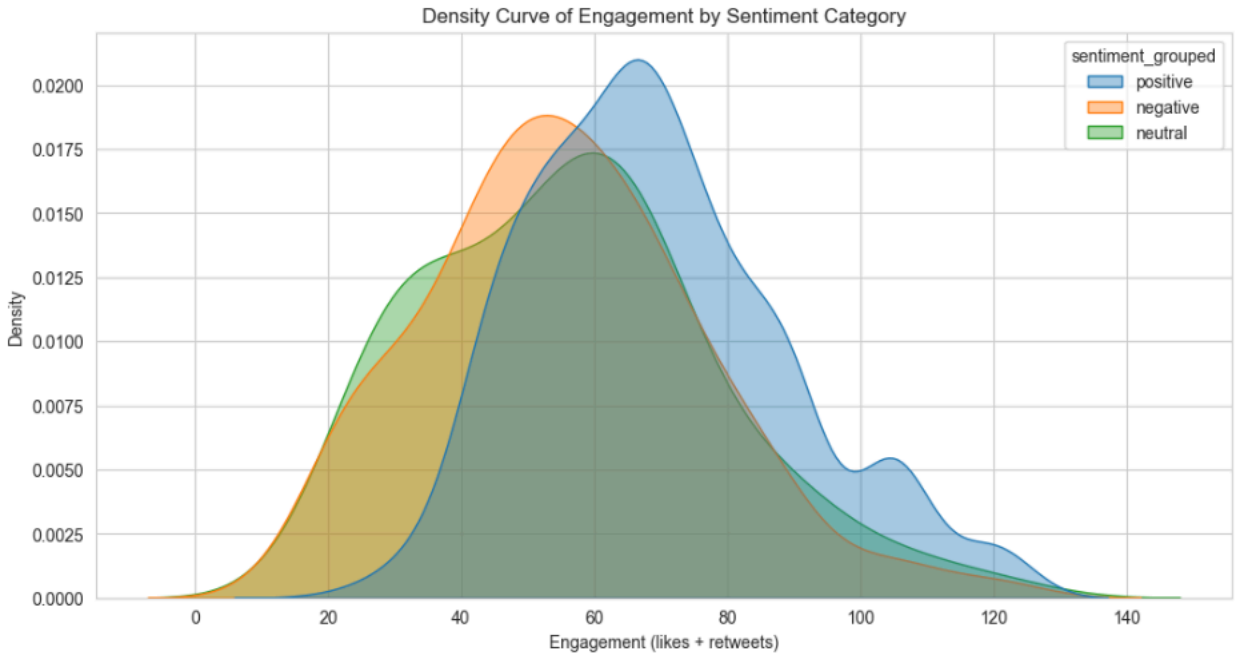


Figure 3. Density Curve of Engagement by Sentiment

Figure 3 shows that **positive posts** have a density curve that extends further to the right, indicating a higher concentration of posts with elevated engagement. **Neutral posts** cluster more toward the lower end of the engagement spectrum, while **negative posts** fall between the two but show considerable overlap.

Also, the visualization suggests that sentiment alone does not strongly separate engagement levels. While the peaks of the distributions differ slightly—suggesting that sentiment may influence typical engagement—each category still contains a wide range of engagement values. This means that highly engaging posts can occur within any sentiment group.

The significant overlap suggests that any differences in engagement by sentiment are likely subtle rather than dramatic. Therefore, visual inspection alone is not sufficient to determine whether these differences are meaningful. This supports the need for regression modeling to formally test whether sentiment has a statistically significant effect on engagement after accounting for variability.

5.2 Sentiment Classifier Performance

The fine-tuned DistilBERT sentiment classifier was evaluated using the held-out test dataset. The model generated predictions for all three sentiment categories: positive, neutral, and negative, and a full classification report was produced. The classifier achieved an overall **accuracy of 0.91**, with a **weighted F1-score of 0.91**, indicating strong performance across the dataset.

To further investigate classification behavior, a confusion matrix was generated.

The confusion matrix below shows how accurately the model distinguished between the three sentiment categories. The diagonal cells represent correct predictions, while off-diagonal cells show misclassifications. The model correctly identified **23 negative**, **6 neutral**, and **66 positive** posts. Most errors occurred within the neutral category, which is expected given the subtle and ambiguous nature of neutral emotional content. Overall, the model demonstrated reliable generalization and produced high-quality sentiment predictions suitable for subsequent engagement analysis.

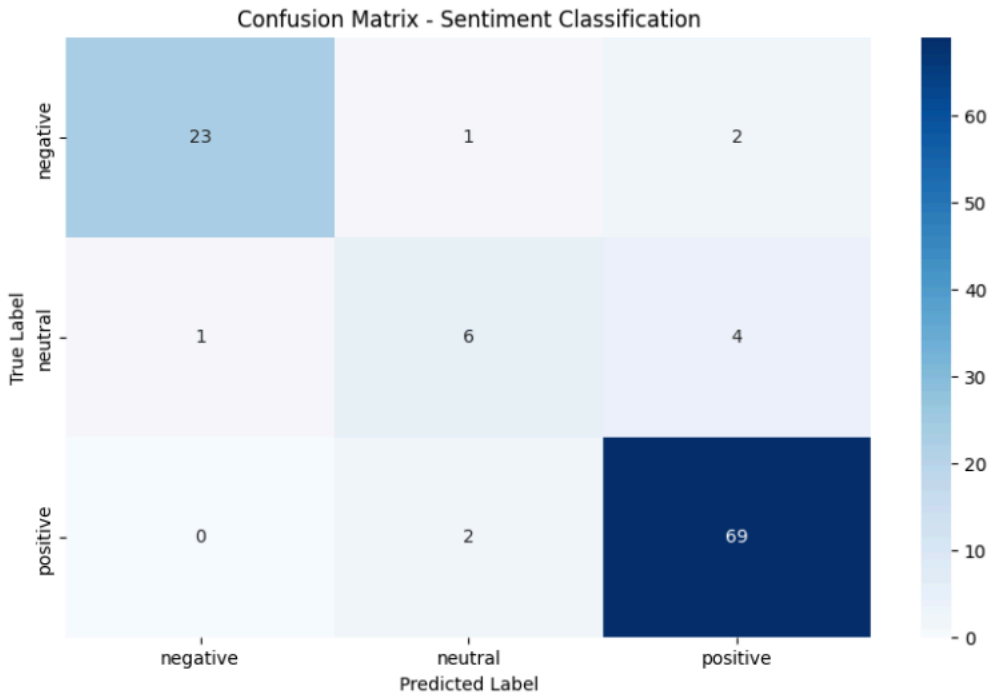


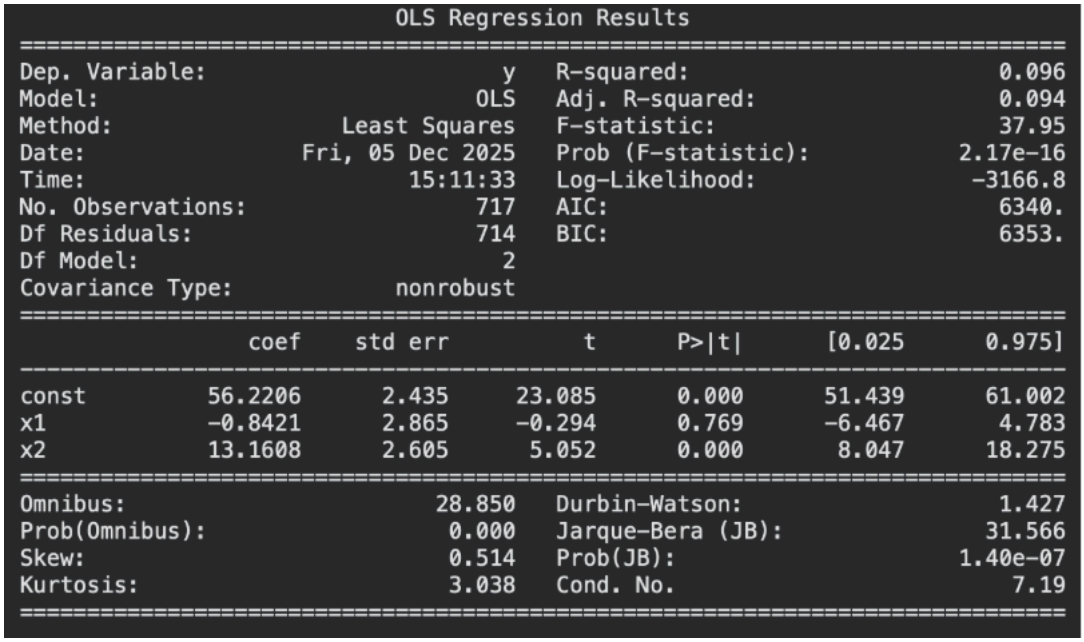
Figure 4. Confusion Matrix for Sentiment Classification

Other Computational Results:

This section presents the analytical outcomes of the sentiment classification model and the statistical evaluation conducted to determine whether sentiment influences social media engagement. Results include classification performance metrics, confusion matrix interpretation, and an Ordinary Least Squares (OLS) regression analysis that examines the relationship between sentiment and engagement.

5.3 Regression Analysis of Sentiment and Engagement

To determine whether sentiment has a measurable effect on user engagement, an Ordinary Least Squares (OLS) regression model was developed using total engagement (likes + retweets) as the response variable. Dummy variables were created for positive and negative sentiment categories, with **neutral** serving as the reference group.



OLS Regression Results

Dep. Variable:	y	R-squared:	0.096
Model:	OLS	Adj. R-squared:	0.094
Method:	Least Squares	F-statistic:	37.95
Date:	Fri, 05 Dec 2025	Prob (F-statistic):	2.17e-16
Time:	15:11:33	Log-Likelihood:	-3166.8
No. Observations:	717	AIC:	6340.
Df Residuals:	714	BIC:	6353.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	56.2206	2.435	23.085	0.000	51.439	61.002
x1	-0.8421	2.865	-0.294	0.769	-6.467	4.783
x2	13.1608	2.605	5.052	0.000	8.047	18.275

Omnibus:	28.850	Durbin-Watson:	1.427
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31.566
Skew:	0.514	Prob(JB):	1.40e-07
Kurtosis:	3.038	Cond. No.	7.19

Table 2. OLS Regression Summary

The regression output shows how engagement levels change for negative and positive posts relative to neutral posts. The model yielded an R^2 value of 0.096, meaning that approximately 10% of the variation in engagement can be explained by sentiment alone. While this indicates that sentiment plays a meaningful role in engagement, the majority of engagement behavior is driven by factors not captured in the model.

Within the regression results:

- The coefficient for negative sentiment was not statistically significant ($p >> .05$), indicating that engagement on negative posts does not differ meaningfully from engagement on neutral posts.
- The coefficient for positive sentiment was statistically significant ($p < .001$), showing that positive posts receive approximately 13 more interactions than neutral posts on average.

These findings indicate that positive emotional tone is associated with significantly higher engagement, while negative sentiment does not systematically increase or decrease user interaction relative to the baseline category.

6. Discussion

6.1 Interpretation of Model Performance

The fine-tuned DistilBERT model demonstrates strong capability in classifying sentiment in social media text. Applying a weighted loss function helped the model handle class imbalances more effectively.

From our analysis, we observed that positive labels were the easiest for the model to learn. This is likely because the positive class was the most prevalent in the dataset, giving the model more examples to generalize from.

In contrast, neutral sentiment proved more challenging to classify. This difficulty can be attributed to several factors:

- Overlap with other classes: Neutral text often shares lexical and semantic features with both positive and negative posts, leading to misclassifications.

- Higher lexical ambiguity: Neutral posts frequently contain vague language, making it harder for the model to assign a definitive label.

- Limited representation: The neutral class had fewer examples in the dataset, which reduces the model's ability to learn robust patterns.
- Grouping strategy challenges: Our approach of grouping sentiments into predefined categories based on labels may have introduced additional difficulty. By forcing posts into these broad categories, subtle distinctions between sentiments could be lost, making it harder for the model to learn nuanced patterns.

Overall, the model performed well on dominant classes (positive and negative) but struggled with underrepresented or semantically ambiguous categories, especially where grouping may have obscured fine-grained sentiment differences.

7. Conclusions

This project demonstrates that a fine-tuned DistilBERT model can reliably classify social media posts into Positive, Negative, and Neutral sentiments. The model performs particularly well on dominant classes, especially positive and negative posts.

Statistical analysis also revealed a significant relationship between sentiment and engagement metrics, with positive content driving the highest likes and retweets. However, overlap in engagement distributions across classes and a low coefficient of determination ($R^2 = 0.096$) indicate that sentiment alone explains only a small portion of engagement variability.

Neutral sentiment remains the most challenging category to classify, due to limited representation and inherent ambiguity. Future work should expand the dataset to include more neutral sentiment examples, incorporate richer engagement predictors, and leverage enhanced modeling approaches to better capture subtle emotional signals. Strengthening these areas will improve model robustness and deepen understanding of the complex drivers behind online engagement.

8. References

"Distilbert/Distilbert-Base-Uncased · Hugging Face." *Huggingface.co*, 11 Mar. 2024, huggingface.co/distilbert/distilbert-base-uncased.

GeeksforGeeks. (2025, March 24). *Distilbert in natural language processing*. https://www.geeksforgeeks.org/nlp/distilbert-in-natural-language-processing/

Distilbert: Multiclass text classification using Transformers(hugging face) | by Preeti | Medium. (n.d.). https://medium.com/@preeti.rana.ai/distilbert-multiclass-text-classification-using-transformers-hugging-face-7c072656525d

AWS. "What Is Sentiment Analysis? - Sentiment Analysis Explained - AWS." *Amazon Web Services, Inc.*, 2023, aws.amazon.com/what-is/sentiment-analysis/.