



Collage of computing

Data science

Data Analysis 2

DS3114

Project report

Task 3: Text Analysis

Instructor: Omaima Fallatah

Student Name	ID
Fajr Faisal Al-zahrani	44410657
Mayar Turki Al-owaydhi	443003550

TABLE OF CONTACT

INTRODUCTION	3
DATASET	4
OBJECTIVES	5
EXPLORATORY DATA ANALYSIS (EDA).....	6
DATA PREPROCESSING.....	10
Split the data.....	15
Naive Bayes.....	16
1.Multinomial Naive Bayes	16
2.Bernoulli Naive Bayes	17
3.Complement Naive Bayes	18
Logistic Regression	20
Comparison Between Naive Bayes and Logistic Regression	21
Conclusion	22

INTRODUCTION

With the rapid growth of social media, vast amounts of text data now reflect public sentiment on a range of topics, making it a valuable resource for understanding trends and opinions. This project aims to develop a sentiment analysis model to automatically classify social media text as positive or negative. Using Natural Language Processing (NLP) for data cleaning, tokenization, and feature extraction, the project prepares the data for accurate classification.

The model will be trained and evaluated on various algorithms, including logistic regression and Naive Bayes, to determine the most effective approach for sentiment prediction. By transforming social media data into meaningful sentiment insights, this model will offer a scalable solution for sentiment analysis in real-time data environments.

DATASET

www.kaggle.com. (n.d.). *Sentiment140 dataset with 1.6 million tweets*.

[online] Available at:

<https://www.kaggle.com/datasets/kazanova/sentiment140>.

OBJECTIVES

1. Develop a Sentiment Classification Model

Build a machine learning model to classify social media text into positive and negative sentiments, using NLP techniques to prepare and process the data effectively.

2. Evaluate Model Performance

Assess the accuracy and reliability of different classification algorithms, such as logistic regression and Naive Bayes, to determine the best approach for sentiment prediction in social media text data.

EXPLORATORY DATA ANALYSIS (EDA)

Description Figure1:

This image displays a portion of a data frame that contains six columns: sentiment, id, date, flag, user, and text.

The data shows individual tweets along with their metadata. The sentiment column contains binary values (0 or 1), where 0 represents a negative sentiment, and the other columns contain the tweet ID, date of the tweet, a flag (set to NO_QUERY), the username, and the tweet text itself.

Purpose of Figure1: The purpose of this figure is to provide a snapshot of the raw tweet data before it is preprocessed. It helps visualize the structure and contents of the dataset, allowing for a better understanding of how tweets are organized and labeled for sentiment analysis.

Figure1

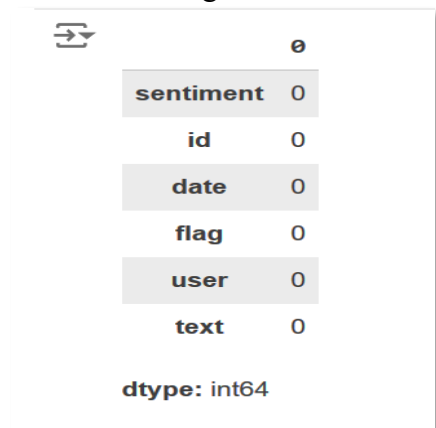


	sentiment	id	date	flag	user	text
0	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
1	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
2	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
3	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all...
4	0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew

Description Figure2: This image shows a summary of the missing values in a pandas DataFrame. Each column (sentiment, id, date, flag, user, and text) has a value of 0, indicating that there are no missing values in any of these columns. The data type displayed is int64.

Purpose of Figure2: The purpose of this figure is to confirm that there are no missing values in the dataset, ensuring that all the data columns are complete and ready for analysis without the need for handling null values.

Figure2



	0
sentiment	0
id	0
date	0
flag	0
user	0
text	0

dtype: int64

Figure3

Description Figure3: shows a summary of a pandas DataFrame with 1,599,999 entries and 6 columns. Each column is listed with its name (sentiment, id, date, flag, user, text), along with the count of non-null values, data types (int64 for numerical data and object for textual data), and memory usage (73.2 MB).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599999 entries, 0 to 1599998
Data columns (total 6 columns):
#   Column      Non-Null Count  Dtype
---  -
0   sentiment    1599999 non-null  int64
1   id           1599999 non-null  int64
2   date         1599999 non-null  object
3   flag         1599999 non-null  object
4   user         1599999 non-null  object
5   text         1599999 non-null  object
dtypes: int64(2), object(4)
memory usage: 73.2+ MB
```

Purpose: The purpose of this figure is to provide an overview of the dataset's structure, including the number of rows and columns, data completeness (no missing values), and the types of data in each column. This is crucial for understanding the dataset before conducting further analysis or preprocessing.

Description of Figure4: shows the count of sentiment labels in the dataset. The sentiment column is divided into two categories: 1 (positive sentiment) and 0 (negative sentiment). There are 800,000 entries labeled as positive (1) and 799,999 entries labeled as negative (0). The data type for the sentiment column is int64.

Figure4

```
count
sentiment
1      800000
0      799999
dtype: int64
```

Purpose: The purpose of this figure is to display the distribution of sentiment labels in the dataset. It provides insight into the balance between positive and negative sentiment data points, which is important for ensuring that models trained on this data are not biased towards one class. The almost equal distribution suggests a balanced dataset.

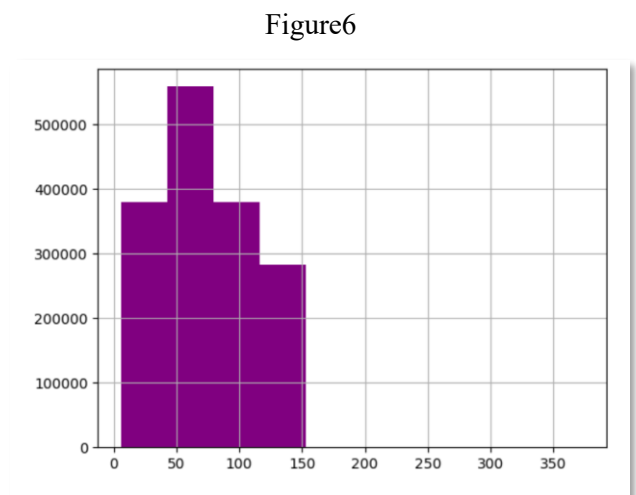
Description Figure5: This bar chart shows the distribution of sentiments in the dataset. The x-axis represents the sentiment labels (0 for negative and 1 for positive), while the y-axis represents the number of tweets. Both bars are nearly equal in height, with around 800,000 tweets for each sentiment class.



The chart uses shades of purple to differentiate between the two classes.

Purpose: The purpose of this figure is to visually depict the balance between positive and negative sentiments in the dataset. It shows that the data is well-balanced, with almost equal numbers of positive and negative tweets, which is important for building a sentiment analysis model that performs fairly across both classes.

Description Figure6: This histogram shows the distribution of the number of characters in each tweet. The x-axis represents the number of characters, while the y-axis represents the frequency of tweets with that character count. Most tweets seem to have between 40 and 100 characters, with a peak around 50 characters. The histogram is colored in dark purple.



Purpose: The purpose of this figure is to visualize the length of tweets in the dataset. It helps in understanding the general distribution of tweet lengths, which can be useful for text preprocessing and feature extraction in sentiment analysis.

Figure7

Description Figure7:

This is a word cloud generated from the text of the tweets in the dataset. The most frequent words appear larger and bolder, while less frequent words appear smaller. Some of the prominent words include "good," "day," "work," "love," "now," and "today," among others.



Purpose: The purpose of this figure is to visualize the most common words used in the tweets. It helps in quickly identifying the key terms that are frequently mentioned in the dataset, which can provide insight into the general topics and sentiments discussed in the tweets.

DATA PREPROCESSING

Description Figure8: This figure shows a part of the final dataset after applying data preprocessing steps to the text. The displayed column is sentiment, where 0 represents negative tweets and 1 represents positive tweets. The values in this column remain unchanged after processing the text column. Several preprocessing operations were applied to the text, including converting all text to lowercase, removing URLs, eliminating special characters and punctuation, and replacing contractions (e.g., changing "can't" to "cannot").

Purpose: The purpose of this figure is to show the state of the data after preprocessing, where the quality of the text has been improved to make it ready for analysis. These steps ensure that the text is standardized and cleaned, making it easier for machine learning algorithms to interpret and analyze. The sentiment column is retained as is, to be used later for training models to classify tweets based on sentiment.

Figure8



sentiment	
0	0
1	0
2	0
3	0
4	0
...	...
1599994	1
1599995	1
1599996	1
1599997	1
1599998	1

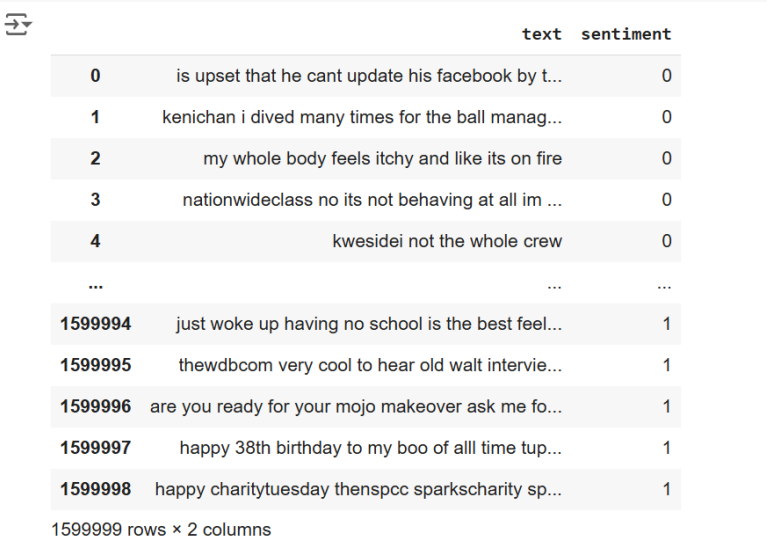
1599999 rows × 1 columns

dtype: int64

Description Figure9: The figure shows the text column containing tweets after being cleaned, along with the sentiment column that classifies the tweets as negative (0) or positive (1). The texts have been cleaned by removing URLs, symbols, and contractions.

Purpose: The purpose is to illustrate how the tweets look after the data preprocessing steps, making the text ready for further analysis and sentiment classification.

Figure9



	text	sentiment
0	is upset that he cant update his facebook by t...	0
1	kenichan i dived many times for the ball manag...	0
2	my whole body feels itchy and like its on fire	0
3	nationwideclass no its not behaving at all im ...	0
4	kwesidei not the whole crew	0
...
1599994	just woke up having no school is the best feel...	1
1599995	thewdbcom very cool to hear old walt intervie...	1
1599996	are you ready for your mojo makeover ask me fo...	1
1599997	happy 38th birthday to my boo of all time tup...	1
1599998	happy charitytuesday thenspcc sparkcharity sp...	1

1599999 rows × 2 columns

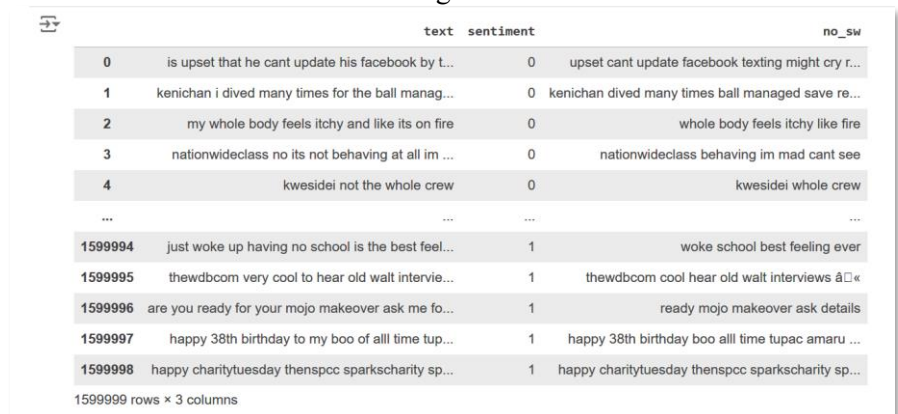
Description

Figure10: The figure displays a data frame with three columns: text (the original tweet text), sentiment (sentiment label, either negative 0 or positive 1), and the last column no_sw, which contains the tweets after

removing stopwords. Stopwords such as "the", "is", and "at" have been removed to retain only the more meaningful words in the text.

Purpose: The purpose of the figure is to show the effect of removing stopwords in the no_sw column. By eliminating these common words, the data becomes less noisy, which improves the accuracy of machine learning models by focusing on the most significant words for sentiment analysis.

Figure10



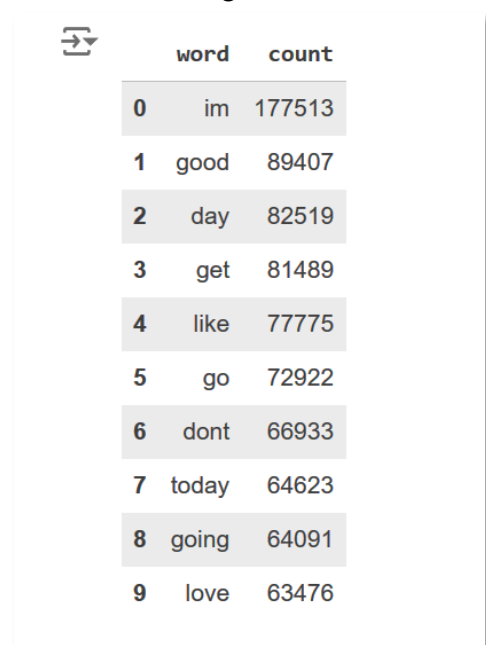
	text	sentiment	no_sw
0	is upset that he cant update his facebook by t...	0	upset cant update facebook texting might cry r...
1	kenichan i dived many times for the ball manag...	0	kenichan dived many times ball managed save re...
2	my whole body feels itchy and like its on fire	0	whole body feels itchy like fire
3	nationwideclass no its not behaving at all im ...	0	nationwideclass behaving im mad cant see
4	kwesidei not the whole crew	0	kwesidei whole crew
...
1599994	just woke up having no school is the best feel...	1	woke school best feeling ever
1599995	thewdbcom very cool to hear old walt intervie...	1	thewdbcom cool hear old walt interviews â«
1599996	are you ready for your mojo makeover ask me fo...	1	ready mojo makeover ask details
1599997	happy 38th birthday to my boo of all time tup...	1	happy 38th birthday boo all time tupac amaru ...
1599998	happy charitytuesday thenspcc sparkscharity sp...	1	happy charitytuesday thenspcc sparkscharity sp...

1599999 rows x 3 columns

Description Figure11: The figure displays a table with two columns: word and count. It shows the ten most frequent words found in the dataset, along with their corresponding counts. The word "im" appears the most, with 177,513 occurrences, followed by words like "good" (89,407 times), "day" (82,519 times), and others like "get", "like", "love", and "go".

Purpose: The purpose of this table is to highlight the most common words in the dataset. This kind of analysis helps in understanding the overall themes and vocabulary commonly used in the tweets. Identifying these frequent words is useful for text preprocessing and feature extraction, as some of these words might be removed in further processing if they don't add value to the sentiment analysis.

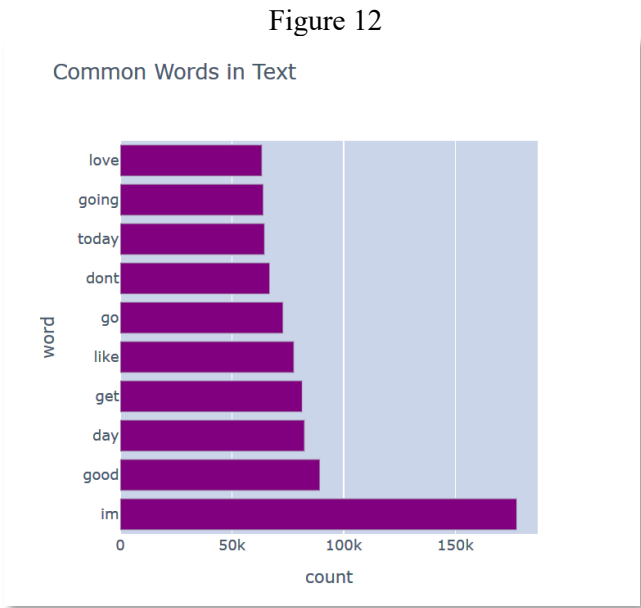
Figure 11



	word	count
0	im	177513
1	good	89407
2	day	82519
3	get	81489
4	like	77775
5	go	72922
6	dont	66933
7	today	64623
8	going	64091
9	love	63476

Description Figure 12: This figure shows a bar chart displaying the most common words found in the text data. The vertical axis lists the words, such as "love," "going," "today," and "im," while the horizontal axis shows the frequency of each word. The word "im" appears the most, with over 150,000 occurrences, followed by words like "good" and "day."

Purpose: The purpose of this bar chart is to provide a visual representation of the most frequent words in the dataset. This helps in understanding the vocabulary that appears most often in the text, which can be important for tasks like sentiment analysis or topic modeling.



Description

Figure 13:

The figure shows a data frame with four columns: text (the original tweets), sentiment (sentiment classification), no_sw (tweets with stopwords removed), and wo_stopfreq (tweets with both stopwords and the most frequent words removed). The wo_stopfreq column focuses on keeping only the most meaningful words by excluding both stopwords and the most common words (like "im", "good", "day") that are frequent but may not add much value to sentiment analysis.

Purpose: The purpose of the wo_stopfreq column is to enhance the quality of the data by removing not only stopwords but also the most frequent words that appear across many tweets. This helps reduce noise and focuses the analysis on the most informative and unique words in the text, improving the performance of machine learning models for sentiment classification.

Figure 13

	text	sentiment	no_sw	wo_stopfreq
0	is upset that he cant update his facebook by t...	0	upset cant update facebook texting might cry r...	upset cant update facebook texting might cry r...
1	kenichan i dived many times for the ball manag...	0	kenichan dived many times ball managed save re...	kenichan dived many times ball managed save re...
2	my whole body feels itchy and like its on fire	0	whole body feels itchy like fire	whole body feels itchy fire
3	nationwideclass no its not behaving at all im ...	0	nationwideclass behaving im mad cant see	nationwideclass behaving mad cant see
4	kwesidei not the whole crew	0	kwesidei whole crew	kwesidei whole crew

Figure 14

	text	sentiment	no_sw	wo_stopfreq	wo_stopfreq_lem
0	is upset that he cant update his facebook by t...	0	upset cant update facebook texting might cry r...	upset cant update facebook texting might cry r...	upset cant update facebook texting might cry r...
1	kenichan i dived many times for the ball manag...	0	kenichan dived many times ball managed save re...	kenichan dived many times ball managed save re...	kenichan dived many times ball managed save re...
2	my whole body feels itchy and like its on fire	0	whole body feels itchy like fire	whole body feels itchy fire	whole body feels itchy fire
3	nationwideclass no its not behaving at all im ...	0	nationwideclass behaving im mad cant see	nationwideclass behaving mad cant see	nationwideclass behaving mad cant see
4	kwesidei not the whole crew	0	kwesidei whole crew	kwesidei whole crew	kwesidei whole crew
...
1599994	just woke up having no school is the best feel...	1	woke school best feeling ever	woke school best feeling ever	woke school best feeling ever
1599995	thewdbcom very cool to hear old walt interview...	1	thewdbcom cool hear old walt interviews â□«	thewdbcom cool hear old walt interviews â□«	thewdbcom cool hear old walt interviews â□«
1599996	are you ready for your mojo makeover ask me fo...	1	ready mojo makeover ask details	ready mojo makeover ask details	ready mojo makeover ask details
1599997	happy 38th birthday to my boo of all time tup...	1	happy 38th birthday boo all time tupac amaru ...	happy 38th birthday boo all time tupac amaru ...	happy 38th birthday boo all time tupac amaru ...
1599998	happy charitytuesday thenspcc sparkscharity sp...	1	happy charitytuesday thenspcc sparkscharity sp...	happy charitytuesday thenspcc sparkscharity sp...	happy charitytuesday thenspcc sparkscharity sp...

1599999 rows × 5 columns

Description Figure 14: The last column `wo_stopfreq_lem`, shows the tweets after applying lemmatization, where words are reduced to their root forms (e.g., "running" becomes "run"). This helps unify different word forms and improves consistency across the text data.

Purpose: The purpose of the `wo_stopfreq_lem` column is to standardize the text by reducing words to their base forms, which enhances model performance by focusing on the most essential form of each word.

Figure 15

	sentiment	text
0	0	upset cant update facebook texting might cry r...
1	0	kenichan dived many times ball managed save re...
2	0	whole body feels itchy fire
3	0	nationwideclass behaving mad cant see
4	0	kwesidei whole crew
...
1599994	1	woke school best feeling ever
1599995	1	thewdbcom cool hear old walt interviews â□«
1599996	1	ready mojo makeover ask details
1599997	1	happy 38th birthday boo all time tupac amaru ...
1599998	1	happy charitytuesday thenspcc sparkscharity sp...

1599999 rows × 2 columns

Purpose: The purpose of this figure is to display the sentiment classification alongside the cleaned tweet text. The text has been preprocessed to make it ready for machine learning tasks, focusing only on the most meaningful words after removing stopwords and frequently occurring terms.

Figure 16

Description figure 16:

The figure shows a data frame with one column labeled `text`, where each tweet is tokenized into a list of words. Each row contains the words from the tweet in list format, maintaining their order but separated into individual tokens.

Purpose: The purpose of this figure is to display the tokenized version of the tweet text. Tokenization is a crucial step in text preprocessing, as it breaks down the text into individual words (tokens), making it easier to analyze and process for machine learning models.

	text
0	[is, upset, that, he, cant, update, his, faceb...
1	[kenichan, i, dived, many, times, for, the, ba...
2	[my, whole, body, feels, itchy, and, like, its...
3	[nationwideclass, no, its, not, behaving, at, ...
4	[kwesidei, not, the, whole, crew]

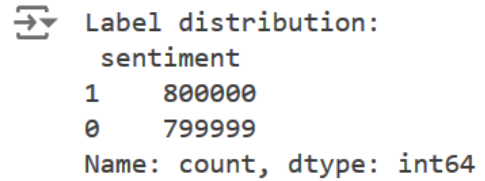
dtype: object

Split the data

1- Label Distribution –Figure 17:

- Displays the count of positive (1) and negative (0) sentiments in the dataset (800,000 positive, 799,999 negative).

Figure 17



```
Label distribution:
sentiment
1      800000
0      799999
Name: count, dtype: int64
```

2- Data Split:

- Splits the data into training (80%) and testing (20%) sets while preserving the class distribution using `train_test_split()`.

3- TF-IDF Vectorization:

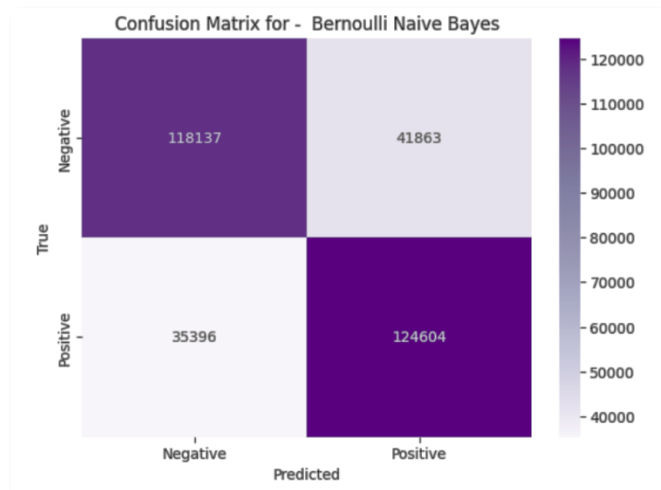
- Converts the text data into numerical features using TF-IDF with a maximum of 5000 features, removing common English stopwords.

2. Bernoulli Naive Bayes

Description: The figure displays the results of a Bernoulli Naive Bayes classifier. The model achieved an accuracy of **75.86%**. It also shows the confusion matrix and classification report:

Confusion Matrix:

- **True Negatives (TN):** 118,137 (correctly predicted negative sentiments)
- **False Positives (FP):** 41,863 (negative sentiments predicted as positive)
- **False Negatives (FN):** 35,396 (positive sentiments predicted as negative)
- **True Positives (TP):** 124,604 (correctly predicted positive sentiments)



Classification Report: It provides precision, recall, F1-score, and support for negative (0) and positive (1) sentiment classes, along with overall accuracy, macro, and weighted averages.

Bernoulli Naive Bayes Accuracy = 75.86%					
Confusion Matrix:					
	0	1			
0	118137	41863			
1	35396	124604			
	precision		recall	f1-score	support
	0	0.77	0.74	0.75	160000
	1	0.75	0.78	0.76	160000
accuracy				0.76	320000
macro avg	0.76	0.76	0.76		320000
weighted avg	0.76	0.76	0.76		320000

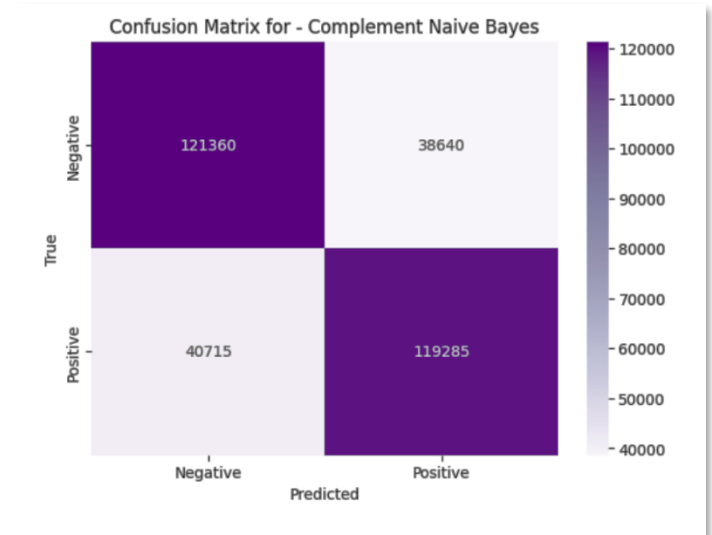
Purpose: The purpose of the confusion matrix and classification report is to visually and quantitatively assess the performance of the Bernoulli Naive Bayes model in predicting sentiment. Together, they help evaluate how accurately the model distinguishes between positive and negative sentiments, using metrics such as accuracy, precision, recall, and F1-score. This comprehensive evaluation provides insights into the model's strengths and areas for improvement.

3.Complement Naive Bayes

Description: The figures display the results of a Complement Naive Bayes classifier. The model achieved an accuracy of **75.20%**. It also shows the confusion matrix and classification report:

Confusion Matrix:

- **True Negatives (TN):**
121,360 (correctly predicted negative sentiments)
- **False Positives (FP):**
38,640 (negative sentiments predicted as positive)
- **False Negatives (FN):**
40,715 (positive sentiments predicted as negative)
- **True Positives (TP):**
119,285 (correctly predicted positive sentiments)



Classification Report: It provides precision, recall, F1-score, and support for negative (0) and positive (1) sentiment classes, along with overall accuracy, macro, and weighted averages.

Complement Naive Bayes Accuracy = 75.20%						
Confusion Matrix:						
	0	1				
0	121360	38640				
1	40715	119285				
			precision	recall	f1-score	support
	0		0.75	0.76	0.75	160000
	1		0.76	0.75	0.75	160000
	accuracy				0.75	320000
	macro avg		0.75	0.75	0.75	320000
	weighted avg		0.75	0.75	0.75	320000

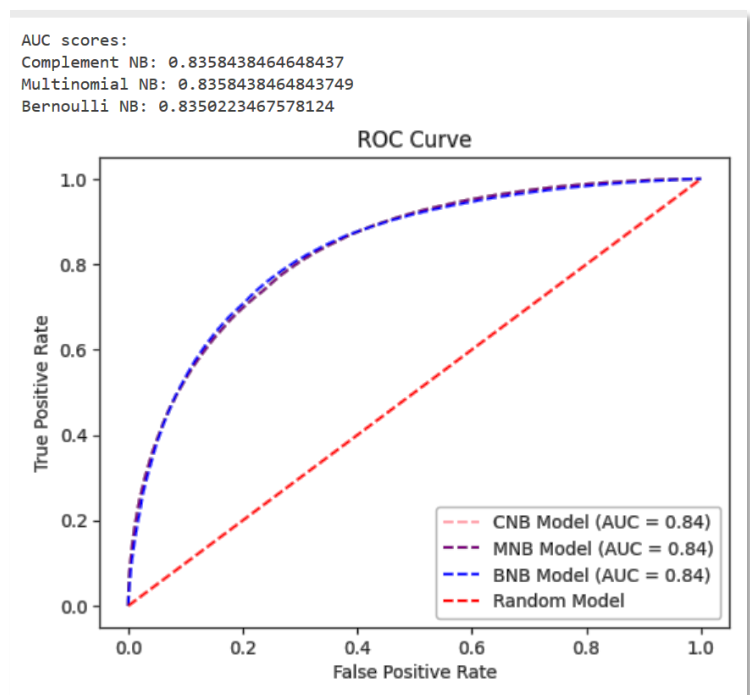
Purpose: The purpose of the confusion matrix and classification report is to visually and quantitatively assess the performance of the Complement Naive Bayes model in predicting sentiment. Together, they help evaluate how accurately the model distinguishes between positive and negative sentiments, using metrics such as accuracy, precision, recall, and F1-score. This comprehensive evaluation provides insights into the model's strengths and areas for improvement.

Description: The image shows the **ROC curve** comparing the performance of three Naive Bayes classifiers: **Complement Naive Bayes (CNB)**, **Multinomial Naive Bayes (MNB)**, and **Bernoulli Naive Bayes (BNB)**, along with a random model as a baseline. The Area Under the Curve (AUC) scores for each model are also provided:

- **Complement Naive Bayes (AUC = 0.84)**
- **Multinomial Naive Bayes (AUC = 0.84)**
- **Bernoulli Naive Bayes (AUC = 0.84)**

The ROC curve plots the **true positive rate (sensitivity)** against the **false positive rate**, which helps evaluate the trade-off between correctly identifying positive instances and misclassifying negatives across different thresholds.

Purpose: The ROC curve and AUC scores are used to visually assess and compare the performance of the classifiers. A higher AUC indicates better model performance in distinguishing between positive and negative classes. The close similarity of the AUC scores for all three Naive Bayes models indicates that they perform similarly in this classification task.



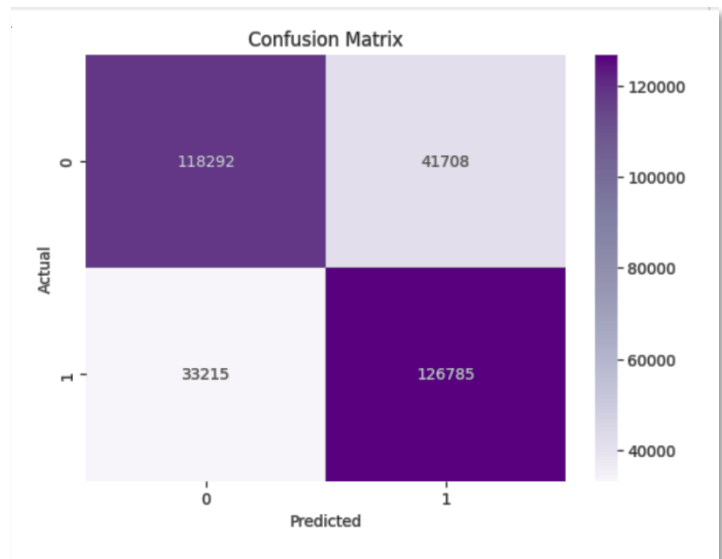
Logistic Regression

Description:

The images display the results of a **Logistic Regression** classifier. The model achieved an accuracy of **77%**. It also shows the confusion matrix and classification report:

Confusion Matrix:

- **True Negatives (TN):** 118,292 (correctly predicted negative sentiments)
- **False Positives (FP):** 41,708 (negative sentiments predicted as positive)
- **False Negatives (FN):** 33,215 (positive sentiments predicted as negative)
- **True Positives (TP):** 126,785 (correctly predicted positive sentiments)



Classification Report: It provides precision, recall, F1-score, and support for negative (0) and positive (1) sentiment classes, along with overall accuracy, macro, and weighted averages.

Logistic Regression Accuracy = 0.77%

Confusion Matrix:

	0	1
0	118292	41708
1	33215	126785

	precision	recall	f1-score	support
0	0.78	0.74	0.76	160000
1	0.75	0.79	0.77	160000
accuracy			0.77	320000
macro avg	0.77	0.77	0.77	320000
weighted avg	0.77	0.77	0.77	320000

Purpose: The purpose of the confusion matrix and classification report is to visually and quantitatively assess the performance of the Logistic Regression model in predicting sentiment. Together, they help evaluate how accurately the model distinguishes between positive and negative sentiments, using metrics such as accuracy, precision, recall, and F1-score. This comprehensive evaluation provides insights into the model's strengths and areas for improvement.

Comparison Between Naive Bayes and Logistic Regression

Accuracy:

- Naive Bayes models achieved around **75.2% to 75.86%**.
- Logistic Regression achieved **77%**, showing better overall accuracy.

Misclassifications:

- Naive Bayes had higher false positives and negatives.
- Logistic Regression had fewer misclassifications, especially in predicting positive sentiments.

Precision, Recall, F1-Score:

- Naive Bayes scores were balanced around **0.75**.
- Logistic Regression had better scores, around **0.77**, making it more accurate in identifying positive sentiments.

Logistic Regression performed better overall in text analysis, with higher accuracy and fewer errors compared to Naive Bayes.

Conclusion

This project demonstrates the potential of machine learning in automating sentiment analysis, transforming raw social media text into actionable insights. By employing NLP techniques to prepare data and training classification models like logistic regression and Naive Bayes, the project successfully classifies text into positive or negative sentiments. The results underscore the effectiveness of combining machine learning with NLP for real-time sentiment analysis, providing a scalable method to gauge public opinion on a large scale. Future work can expand upon this model by incorporating additional data sources, refining algorithms, or exploring deep learning techniques to enhance prediction accuracy and adapt to evolving language patterns on social media platforms.