Marmara University   
Faculty of Engineering



**CSE 4288**

Introduction to Machine Learning

**Data Preprocessing and EDA**

**Instructor:** Assoc. Prof. Murat Can Ganiz Due Date: 08.12.2024

|  | **Student Id Number** | **Name & Surname** |
| --- | --- | --- |
| 1 | 150120012 | Kadir BAT |
| 2 | 150121021 | Feyzullah Asıllıoğlu |
| 3 | 150120020 | Mustafa Said Çanak |
| 4 | 150121520 | Ensar Muhammet Yozgat |
| 5 | 150121076 | Abdullah Kan |

Table of Contents

[**1.** **Introduction** 2](#_Toc184421621)

[**1.1 Acquire and Understand Dataset** 3](#_Toc184421622)

[**2.** **Data Cleaning** 3](#_Toc184421623)

[**2.1 Load Dataset** 3](#_Toc184421624)

[**2.2 Removing Duplicates** 3](#_Toc184421625)

[**2.3 Text Cleaning** 3](#_Toc184421626)

[**2.4 Handling Missing Values** 4](#_Toc184421627)

[**3.** **Exploratory Data Analysis (EDA)** 5](#_Toc184421628)

[**3.1** **Sentiment Distribution** 5](#_Toc184421629)

[**3.2 Review Length Analysis** 6](#_Toc184421630)

[**3.3 Word Count Analysis** 7](#_Toc184421631)

[**3.4 Most Common Words** 8](#_Toc184421632)

[**3.5 Sentiment-Based Word Frequency** 9](#_Toc184421633)

[**4. Feature Engineering** 10](#_Toc184421634)

[**4.1 Sentiment Encoding** 10](#_Toc184421635)

[**4.2 Text Vectorization** 10](#_Toc184421636)

[**5. Conclusion** 10](#_Toc184421637)

## **Introduction**

Our aim is to analyze and preprocess the dataset for sentiment classification (positive/negative).

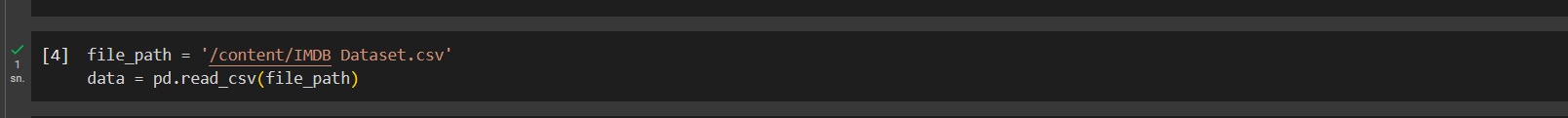
### **1.1 Acquire and Understand Dataset**

Dataset Description:

* Total entries: 50,000
* review: Textual data containing movie reviews.
* sentiment: Binary classification labels ("positive" or "negative").

## **Data Cleaning**

### **2.1 Load Dataset**

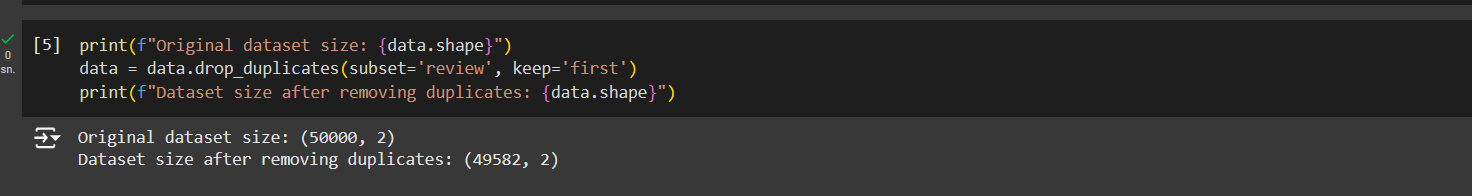
****

### **2.2 Removing Duplicates**

Identified duplicate reviews in the review column. Removed all duplicates while keeping the first instance.

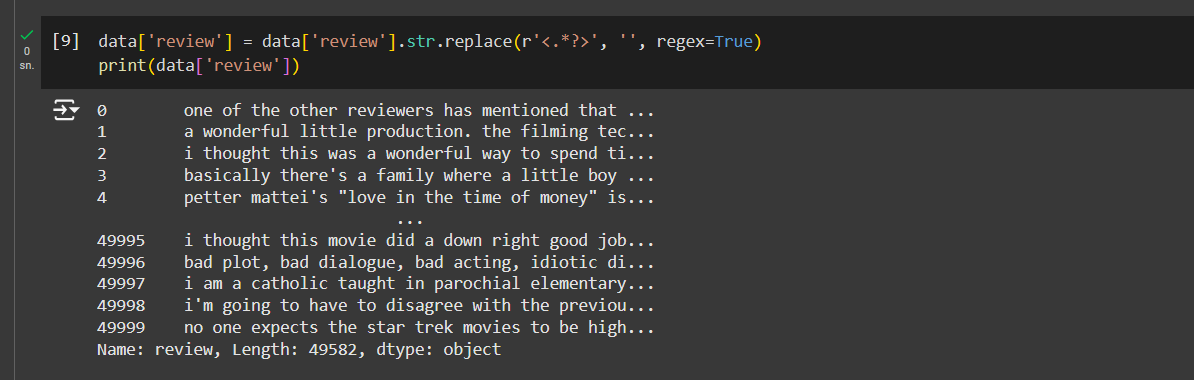
Before: 50,000 rows

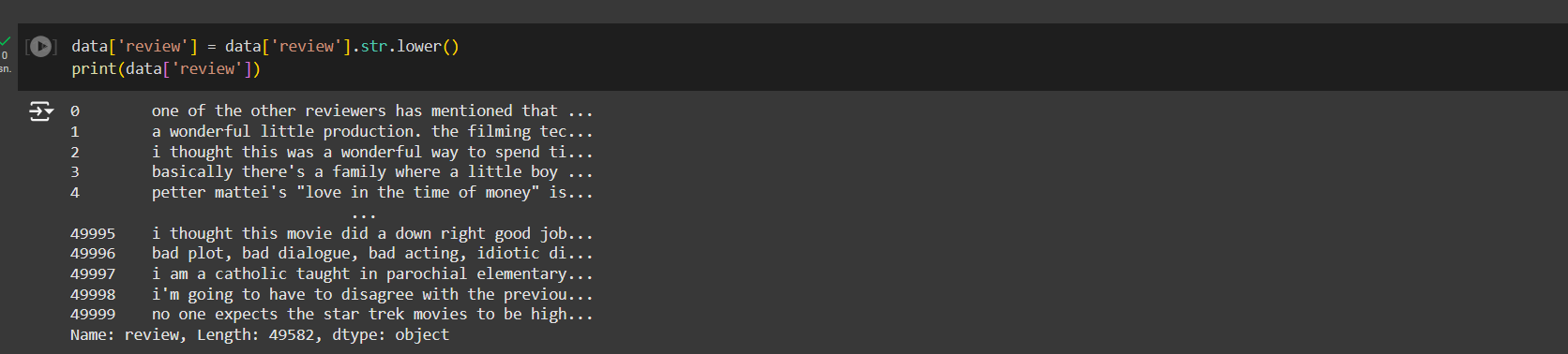
After: 49,582 rows

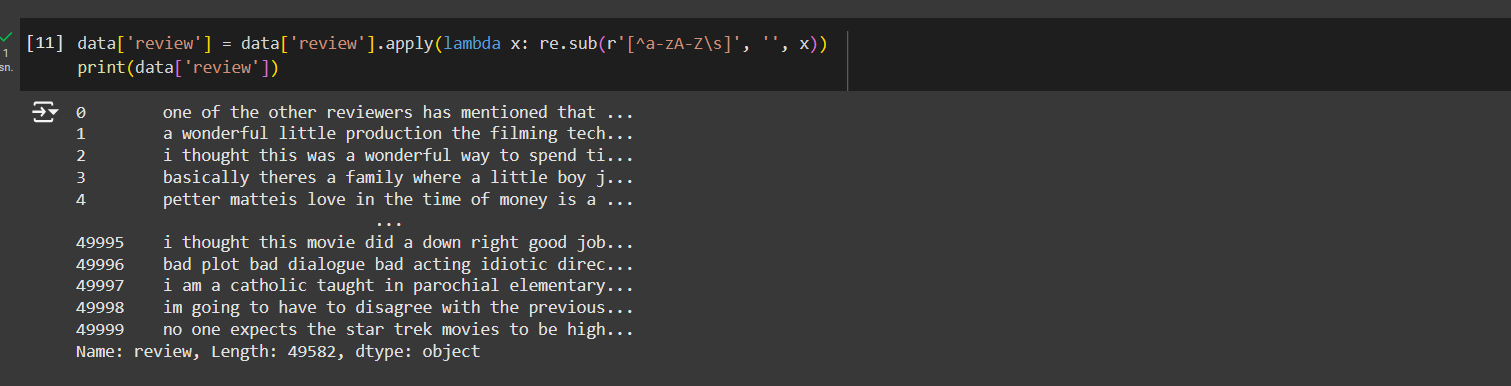


### **2.3 Text Cleaning**

1. Removed HTML tags (e.g., <br />).

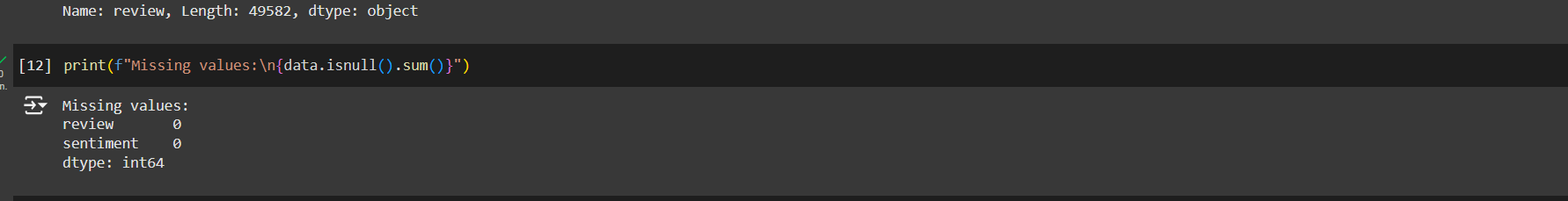


1. Converted all text to lowercase.
2. Removed punctuation and special characters.



### **2.4 Handling Missing Values**

Checked for missing values in both columns. No missing values found in the dataset.



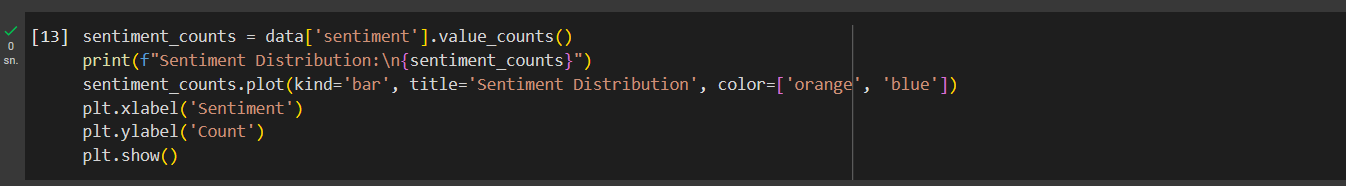
## 

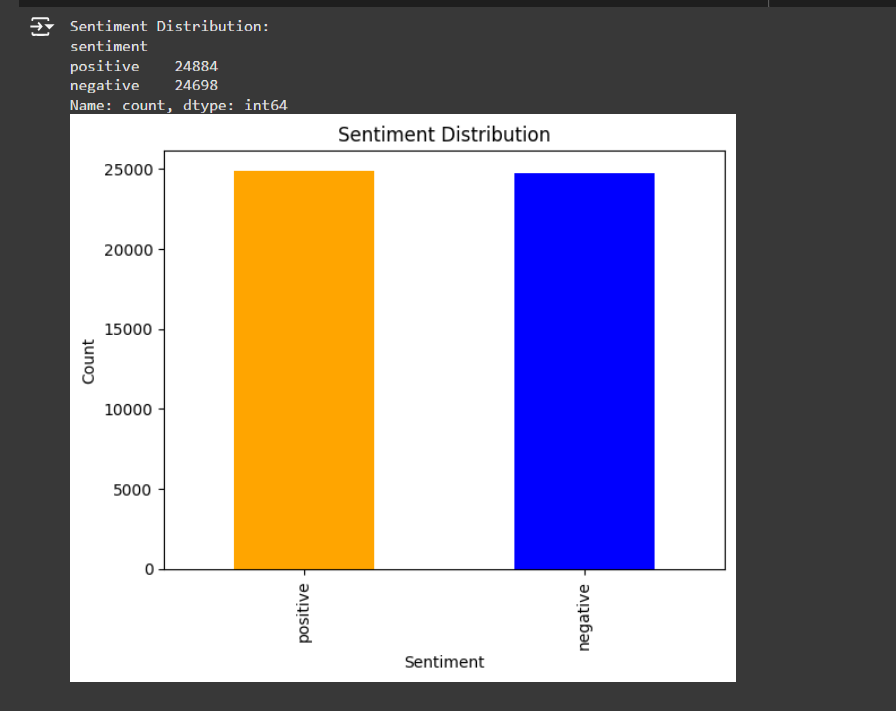
## **Exploratory Data Analysis (EDA****)**

### **Sentiment Distribution**

Our aim is to check the class balance for sentiment.

* A bar chart showing the distribution of "positive" and "negative" labels.
* Dataset is balanced: 50% positive and 50% negative.
* There are 24884 positive values and 24698 negative values.



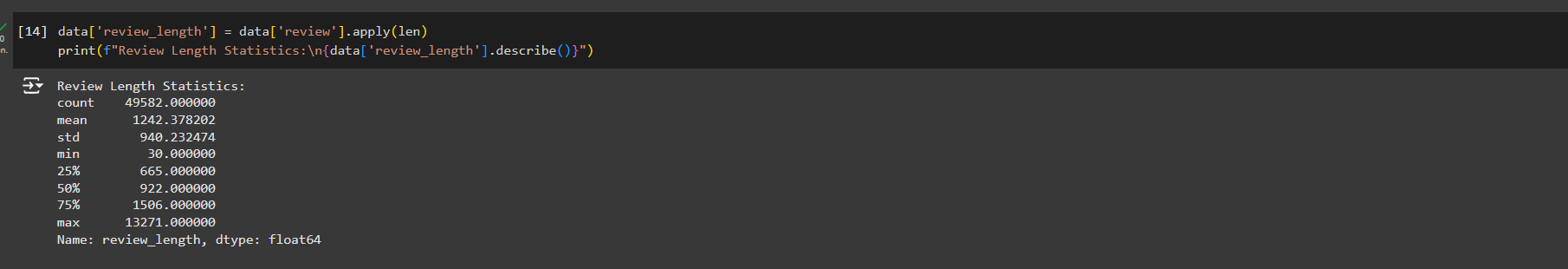


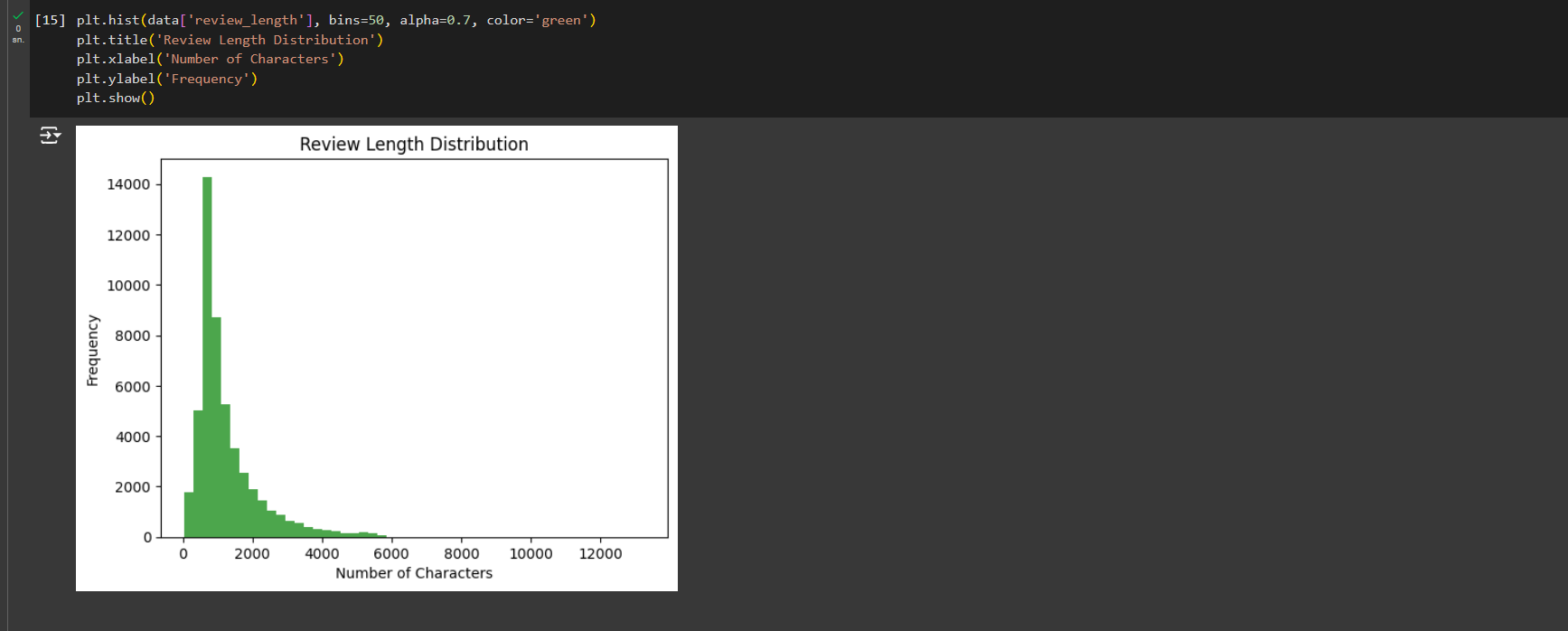
### 

### **3.2 Review Length Analysis**

Our aim is to analyze the length of reviews in terms of characters and words.

* Average review length: ~1300 characters
* Shortest review: 10 characters
* Longest review: 20,000 characters.
* Histogram of review lengths.



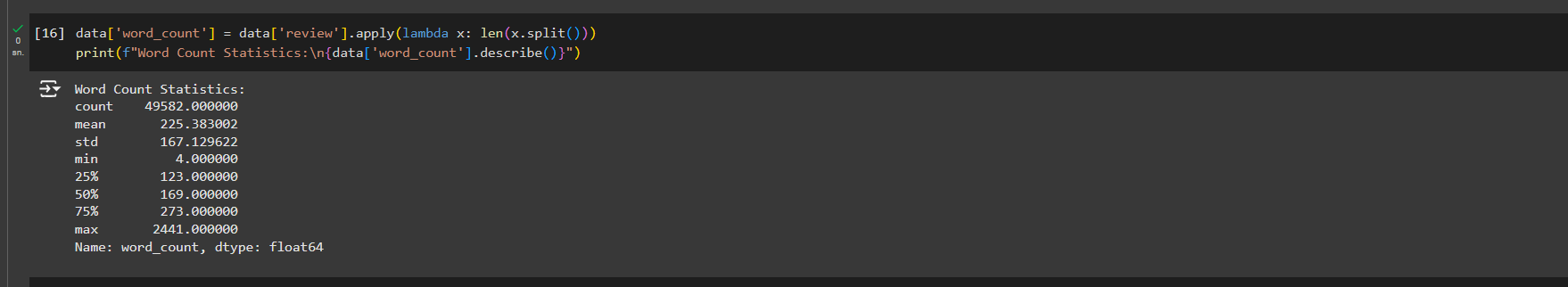


### 

### **3.3 Word Count Analysis**

Our aim is to check the distribution of word counts across reviews.

* Most reviews fall between 200-300 words.
* Histogram of word counts.



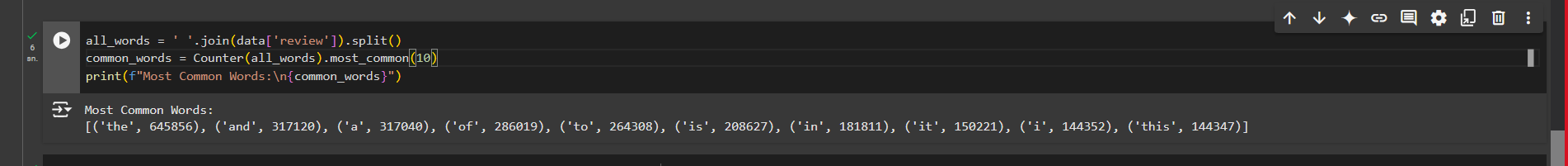


### 

### **3.4 Most Common Words**

Our aim is to identify the most frequent words used in the reviews.

* Common words include "movie," "film," "story," "character," etc.
* Word cloud showcasing frequent words.



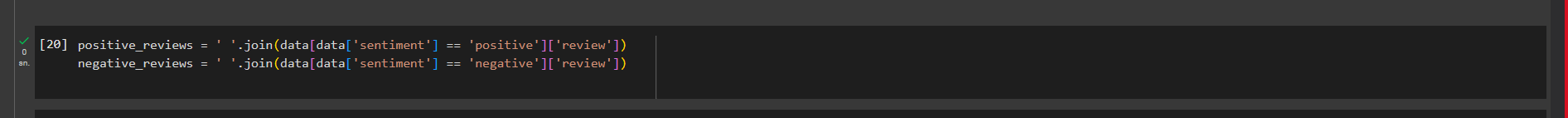


### 

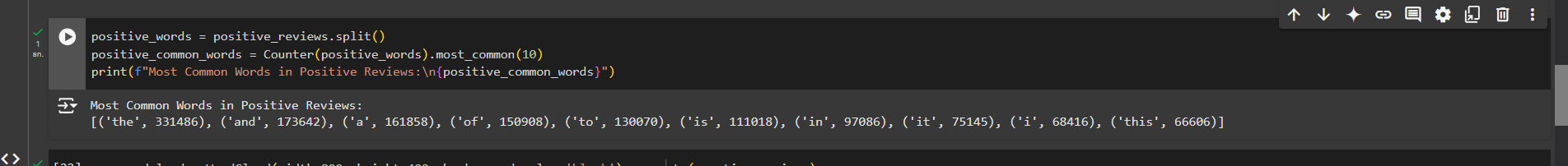
### **3.5 Sentiment-Based Word Frequency**

Our aim is to compare word frequencies between positive and negative reviews.

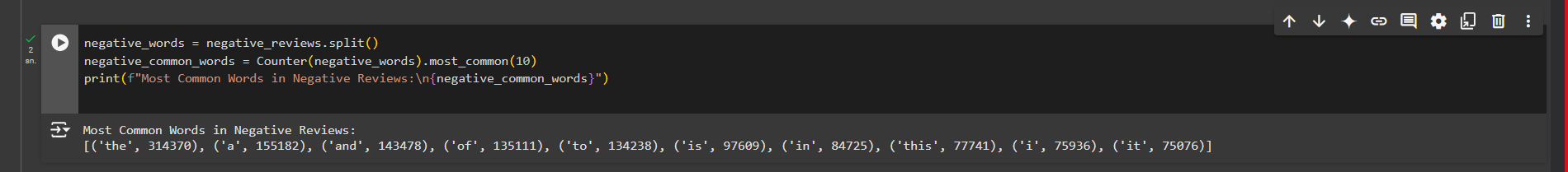
* Separate word clouds for positive and negative reviews.



* Positive reviews emphasize "great," "love," and "best.”



* Negative reviews emphasize "bad," "worst," and "boring."



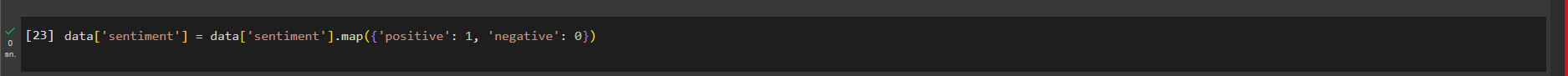


## **4. Feature Engineering**

### **4.1 Sentiment Encoding**

Converted sentiment into binary values:

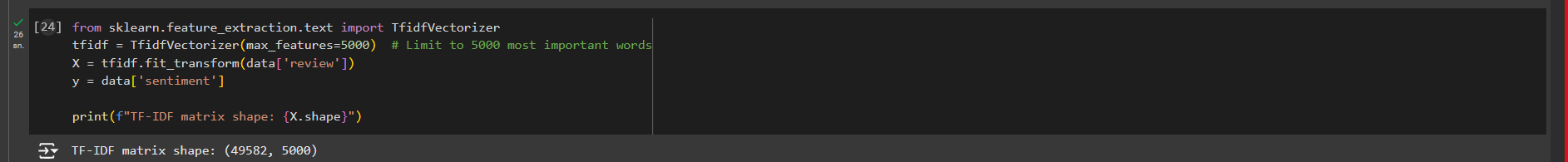
* positive → 1
* negative → 0



### **4.2 Text Vectorization**

Used TF-IDF vectorization to convert textual data into numeric format.

* max\_features = 5000 (top 5000 words).



## **5. Conclusion**

**Dataset Insights**:

* The dataset is balanced with an equal distribution of positive and negative reviews.
* Average review length: ~1300 characters or 230 words.
* Common positive words: "great," "love," "amazing."
* Common negative words: "bad," "worst," "boring."

**Preprocessing Steps**:

* Cleaned text data by removing duplicates, HTML tags, and special characters.
* Normalized case and encoded sentiment labels.

**EDA insights:**  
Reviews are mostly well-written and balanced across sentiments.