Data mining HW2: Report

Malik DAHMANI

May 14, 2024

Contents

1 Question 1: How do you select features for your model input, and what preprocessing did you perform to review text?

For data pre-processing, I first combined the title and text tags into a single column. By increasing the element to tokenize, I will increase the amount of text data available for tokenization, which improved the performance of my model. Moreover, the rating tag are converted from 1-5 to 0-4.

For the data tokenisation, I used of a pre-trained model available on HuggingFace, specifically the "LiYuan/amazon-review-sentiment-analysis" model. This model provides a tokenisation adapted to Amazon review sentiment analysis. During the tokenisation process, if the text of an example exceeds a predefined maximum length of 30 tokens, it is truncated to ensure a uniform length for each entry. At the same time, if the text is shorter, padding is applied to fill in the missing tokens until the specified maximum length is reached.

Finally, I repeated these steps for the test data to make predictions. This meant combining the text, tokenising with truncation and padding, and converting the data into a format compatible with the model. Once the data was pre-processed, I used the trained model to predict the corresponding labels for each test example.

2 Question 2: Please describe how you tokenize your data, calculate the distribution of tokenized sequence length of the dataset and explain how you determine the padding size

Data tokenization is performed in the ”preprocess_function”. I use the tokenized pre-trained model ”LiYuan/amazon-review-sentiment-analysis”. The parameters used for tokenization are: truncation = True, padding =’max_length’, max_length =30). .

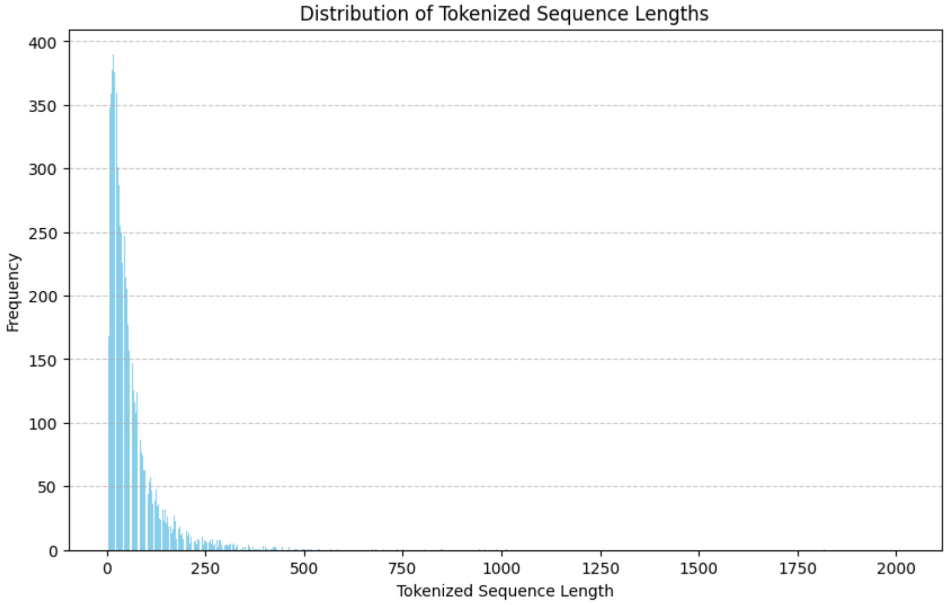

Figure 1: Figure showing the distribution of tokenized sequence length

To choose the padding size, I used several metrics to evaluate the distribution of tokenised sequence lengths in the dataset. In particular, I calculated the mean, the median and the token length with the highest frequency.

The token length with the most sequences is 17, indicating that the majority of examples in the dataset have a tokenised sequence length of 17 tokens. The average tokenised sequence length is around 51, but this is influenced by the longest sequences. The median tokenised sequence length is 8, indicating that half of the examples have a tokenised sequence length of less than 8 tokens.

Taking these metrics into account, and to ensure an appropriate sequence size for the model while avoiding excessive computational cost, I chose a padding size of 30, which lies between the mean and the median. This value offers a reasonable compromise between considering shorter sequences and handling longer sequences without overloading the model with an excessive number of tokens.

3 Question 3: Please compare the impact of using different methods to prepare data for different rating categories

To answer this question, I made three predictions. First, I trained three models:

- 1st model: title and text
- 2nd model: title only
- 3rd model: text only

Due to limited computational resources, I then made predictions with each model on 500 examples from each class and plotted the results.

The second plot,where predictions were made using both title and text, shows a more balanced distribution across most classes, with a slight drop in the 4-star class.

The third plot, showing predictions based on the title only, indicates a skew towards the 3-star class, suggesting that titles alone may not be as indicative of the class as the full text.

Finally, the fourth plot, where predictions were made using only text, there is a noticeable distribution with the highest count in the 1-star, 3-star and 5-star classes.

The results indicate that using both title and text together provides a more balanced prediction across the classes. Predictions based only on the title are skewed towards the 3-star class, suggesting that titles alone may not sufficiently represent the overall sentiment. Predictions made using only the text result in higher counts for the 1-star, 3-star, and 5-star classes, indicating that text alone can capture a larger range of sentiments but may lead to more pronounced peaks in certain classes. Therefore, combining title and text appears to yield the most balanced and accurate classification.

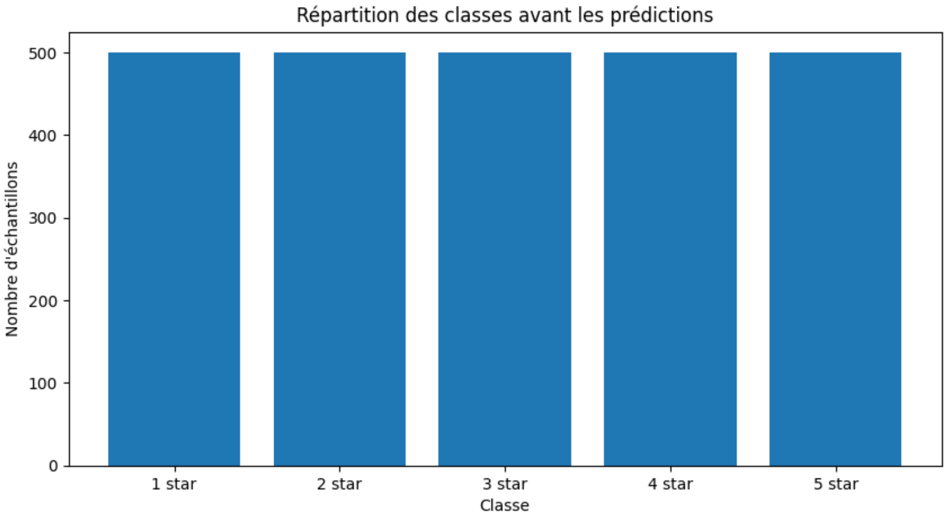

Figure 2: Figure showing the distribution of the examples

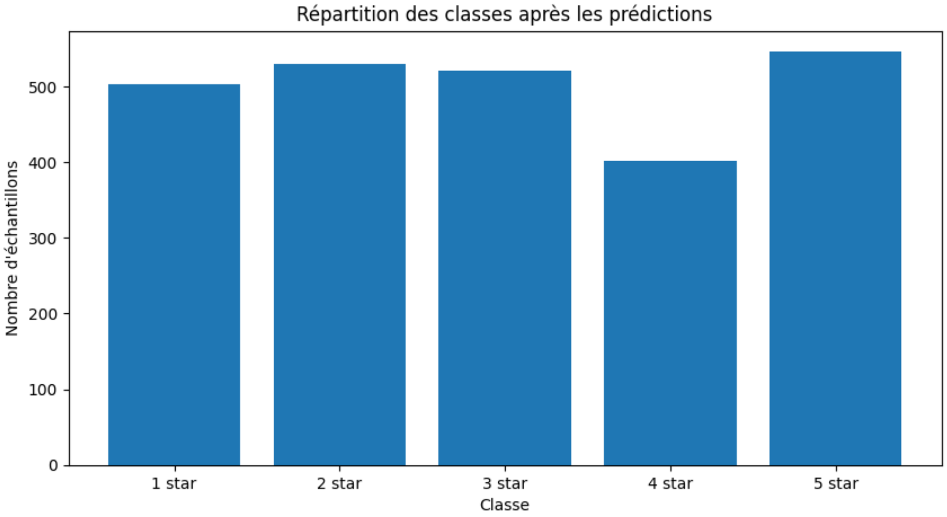

Figure 3: Figure showing the distribution of the examples using text and title

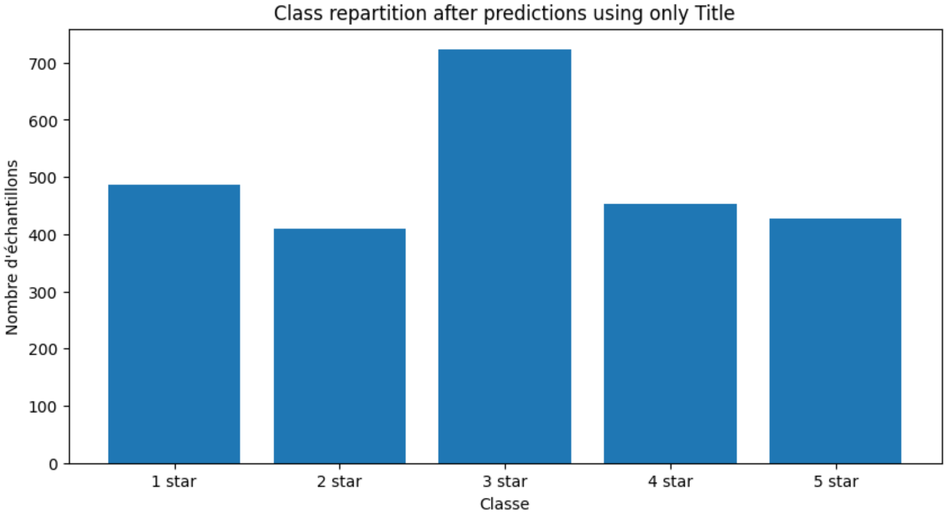

Figure 4: Figure showing the distribution of the examples using only title

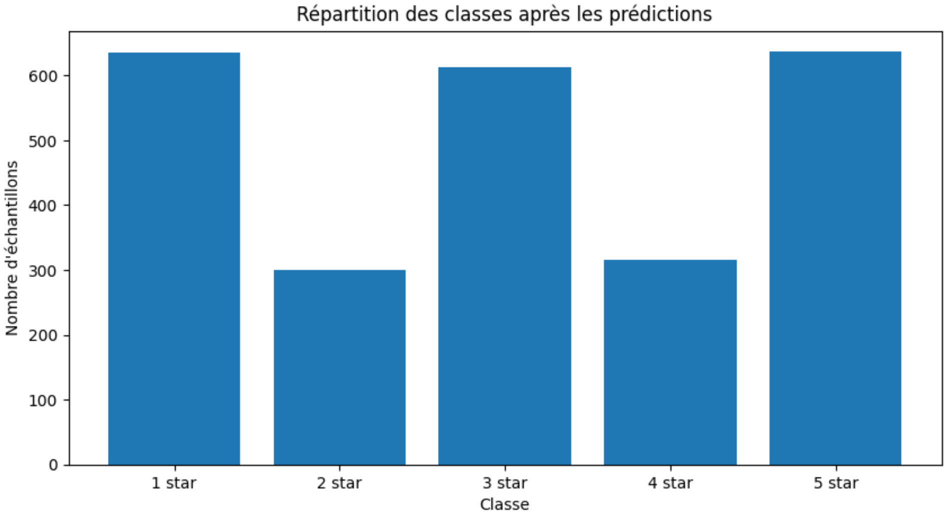

Figure 5: Figure showing the distribution of the examples using only text