HLT Project Proposal

In this brief report, our aim is to introduce our team and provide an overview of the work we will be carrying out for the HLT project.

Group

Let's begin by saying that our group ID is 11. The group consists of three individuals, whose information we will present in the table below:

First Name	Last Name	E-mail	Degree Course
Angelo	Nardone	a.nardone5@studenti.unipi.it	AI
Riccardo	Marcaccio	r.marcaccio@studenti.unipi.it	AI
Matteo	Ziboli	m.ziboli@studenti.unipi.it	Linguistics

Motivation

We believe that understanding whether a product review is positive or negative is useful for several reasons. Firstly, such a system can enhance user experience by providing insights and recommendations on the quality of the product in question. Additionally, businesses can also gain valuable insights from customer feedback, enabling them to make informed decisions to improve their products. Lastly, an NLP system for sentiment analysis can help companies manage their online reputation more effectively by promptly addressing negative feedback and fostering positive engagement with customers.

Goal of the Project

The purpose of our project is to take as input product reviews collected from Amazon (we will discuss the dataset later) and be able to distinguish between positive and negative reviews.

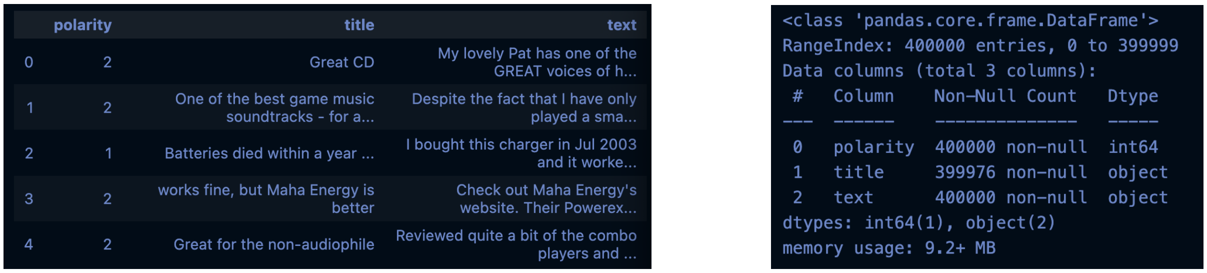
Our idea is to attempt binary classification using only the review titles as input, such as ”`Great CD`” or ”`Batteries died within a year`”. Clearly, titles are usually short and impactful phrases. Therefore, we believe that these may be sufficient to effectively operate our classifier. However, we plan to extend this idea by also attempting classification using the entire reviews as input. Using the entire reviews entails processing longer sentences and greater computational effort, but it also provides more information and potentially higher accuracy. At this point, our plan is to compare the results of these two classifiers using various metrics to determine which approach is more effective.

Upon the recommendation of our professor, we are including an additional potential task to implement. If the above mentioned work proves to be too simple, we have identified a second dataset (which we will discuss shortly) that again would allow us to input product reviews from Amazon and classify them as positive or negative reviews. In this case, we could evaluate our algorithms on two different datasets to assess their effectiveness. Additionally, this second dataset would not only enable binary classification (positive or negative review), but also allow us to attempt to predict the score associated with the review (ranging from 1 for very negative reviews to 5 for very positive reviews). In this scenario, we can view it as a multi-class classification task, which would certainly make our work more interesting from a didactic perspective.

Available Data

Below, we present the two datasets we have identified, noting that the first dataset is the one we will primarily work on, while the second one will be used in case the workload on the first dataset proves to be too simple.

Dataset 1 - Amazon Reviews



	polarity	title	text
0	2	Great CD	My lovely Pat has one of the GREAT voices of h...
1	2	One of the best game music soundtracks - for a...	Despite the fact that I have only played a sma...
2	1	Batteries died within a year ...	I bought this charger in Jul 2003 and it worke...
3	2	works fine, but Maha Energy is better	Check out Maha Energy's website. Their Powerex...
4	2	Great for the non-audiophile	Reviewed quite a bit of the combo players and ...

(a)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400000 entries, 0 to 399999
Data columns (total 3 columns):
 #   Column    Non-Null Count   Dtype 
---  ------    --------------   ----- 
 0   polarity  400000 non-null  int64 
 1   title     399976 non-null  object
 2   text      400000 non-null  object
dtypes: int64(1), object(2)
memory usage: 9.2+ MB
```

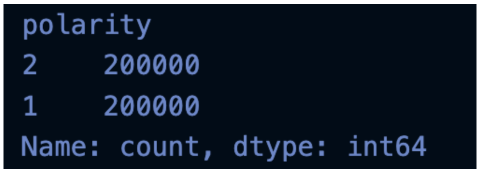
(b)

Figure 1: (a) An example of the first 5 columns of the dataset printed with the command `.head()` of pandas. (b) Information printed with the command `.info()` of pandas.

Here is some useful information about our dataset:

- **Available at:** Amazon Reviews Dataset.
- **Context:** The Amazon reviews dataset consists of reviews from Amazon. The data span a period of 18 years, including 400.000 reviews up to March 2013. The Amazon Reviews dataset is constructed by Xiang Zhang. For more information: [1].
- **Content:** The dataset contains 3 different columns: polarity, title, text. Here's a brief description of these features:

  1. **Polarity:** In this column, a number between 1 and 2 is written. In constructing the dataset, the label 1 was inserted for all negative reviews (those with 1 or 2 stars), while the label 2 was inserted for all positive reviews (4 or 5 stars). Reviews with 3 stars were ignored. The dataset contains 200,000 data points for each class.

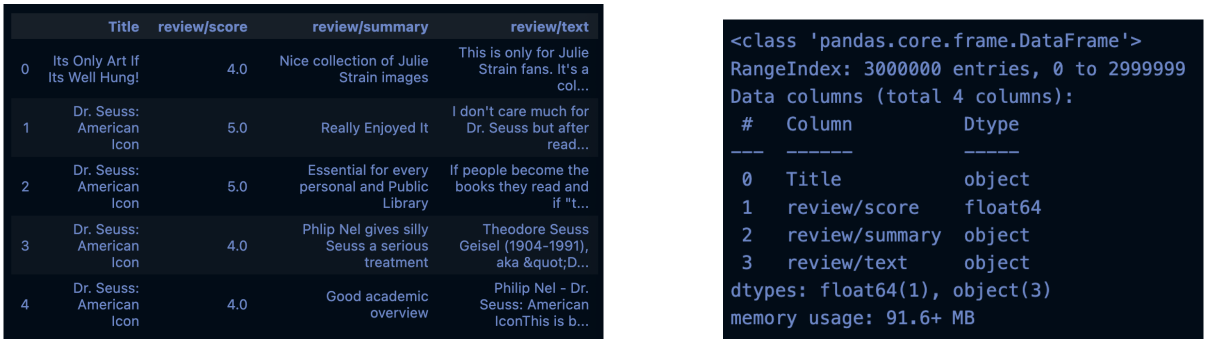


```
polarity
2    200000
1    200000
Name: count, dtype: int64
```

Figure 2: The distribution of polarity values in the dataset printed using the command `.value_counts()` of pandas.

2. **Title:** It contains the titles of the reviews.
3. **Text:** It contains the entire text of each review.

Dataset 2 - Amazon Books Reviews



	Title	review/score	review/summary	review/text
0	Its Only Art If Its Well Hung!	4.0	Nice collection of Julie Strain images	This is only for Julie Strain fans. It's a col...
1	Dr. Seuss: American Icon	5.0	Really Enjoyed It	I don't care much for Dr. Seuss but after read...
2	Dr. Seuss: American Icon	5.0	Essential for every personal and Public Library	If people become the books they read and if "t...
3	Dr. Seuss: American Icon	4.0	Phlip Nel gives silly Seuss a serious treatment	Theodore Seuss Geisel (1904-1991), aka &quot;D...
4	Dr. Seuss: American Icon	4.0	Good academic overview	Philip Nel - Dr. Seuss: American IconThis is b...

(a)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000000 entries, 0 to 2999999
Data columns (total 4 columns):
 #   Column          Dtype
---  ------          -----
 0   Title           object
 1   review/score    float64
 2   review/summary  object
 3   review/text     object
dtypes: float64(1), object(3)
memory usage: 91.6+ MB
```

(b)

Figure 3: (a) An example of the first 5 columns of the dataset printed with the command `.head()` of pandas. (b) Information printed with the command `.info()` of pandas.

Here is some useful information about our dataset:

- **Available at:** Amazon Books Reviews Dataset.
- **Context:** The dataset contains feedback from approximately 3 million users on 212,404 unique books, covering product reviews spanning from May 1996 to July 2014. This dataset was compiled from the Amazon website. For more information, the dataset was created as part of the following research works: [2], [3].
- **Content:** The dataset contains a lot of columns. For our purpose, we will focus only on three columns: title, review/score, review/summary, review/text. Here's a brief description of these features:
  1. **Title:** It contains the book titles on which the review is based.
  2. **Review/Score:** It contains the score, from 1 to 5 stars, of each review. Specifically, 1 star indicates that the review is strongly negative, while 5 stars indicate that the review is strongly positive. The distribution of classes is shown in the Figure 4.

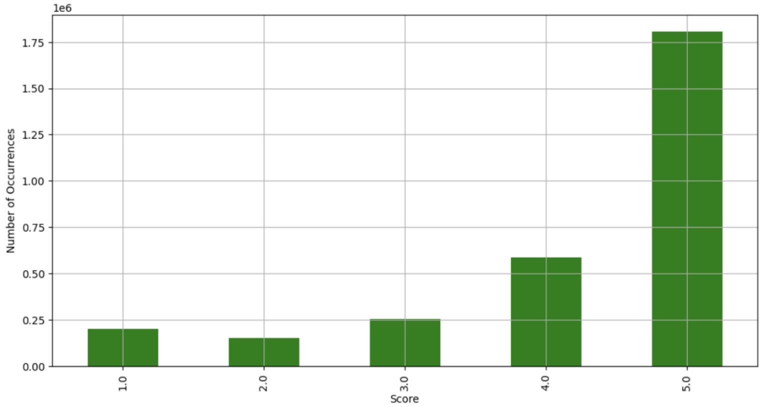


Figure 4: The distribution of review/score values in the dataset.

3. **Review/Summary:** It contains the titles of the reviews.
4. **Review/Text:** It contains the entire text of each review.

Implementation and Evaluation

Let's briefly discuss how we plan to implement our project. Clearly, we haven't covered many topics in the course yet, so this is our preliminary idea. To achieve our goal of binary classification, we intend to primarily use transformers. Since our main task is sentiment analysis, we thought of complementing these transformers with lexicons. For instance, lexicons can be used as additional features to enrich the text representation before passing it through the transformer model. Alternatively, polarity scores assigned by the lexicons can be used as additional labels during the pre-training of the transformer model, such as in training a regression model to predict polarity scores. However, we understand that these are aspects we will explore further as we progress.

Additionally, we will employ various evaluation techniques. To assess the performance of our model, we plan to compute several metrics such as accuracy, precision, recall, F1 score, confusion matrix, and ROC curve. Moreover, we will not only evaluate the effectiveness of our model but also compare its results with different inputs (e.g., titles and full reviews). For this type of analysis, we will also consider the training time and the amount of space used to train our model as evaluation metrics.

We acknowledge that these plans are subject to refinement as we delve deeper into the project and acquire more knowledge during the course.

References

[1] J. McAuley and J. Leskovec. *Hidden factors and hidden topics: understanding rating dimensions with review text* RecSys, 2013.

[2] R. He, J. McAuley. *Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering.* WWW, 2016.

[3] J. McAuley, C. Targett, J. Shi, A. van den Hengel. *Image-based recommendations on styles and substitutes.* SIGIR, 2015.