HLT Project: Progress Update

Sentiment Analysis on Amazon Reviews

HLT - Group 11

Angelo Nardone, Riccardo Marcaccio, Matteo Ziboli

April 18, 2024

Dipartimento di Informatica
Università di Pisa

Introduction

HLT Project: Progress Update

- We already discussed the two datasets we'll be utilizing for our project and outlined our objectives.
- We've delved deeper into our datasets.
- We will now see how we handled the following parts:
  1. Data Loading
  2. Data Cleaning
  3. Data Analysis
- The next step is to start working on our classification task using machine learning.

Table of contents

Using Pandas

HLT Project: Data Loading

- Used `Pandas` to handle the data.
- Loaded both datasets using the command `pd.read_csv()`.
- Quickly saw the contents of the datasets.

```
Dataset_1 = pd.read_csv("Datasets/Dataset_1/train.csv", header=None, names=["polarity", "title", "text"])
```

The dataset contains 3 different columns: *polarity*, *title*, *text*. Here's a brief description of these features:

- *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.
- *title*: It contains the titles of the reviews.
- *text*: It contains the entire text of each review.

```
Dataset_1.head()
```

	polarity	title	text
0	2	Stuning even for the non-gamer	This sound track was beautiful! It paints the ...
1	2	The best soundtrack ever to anything.	I'm reading a lot of reviews saying that this ...
2	2	Amazing!	This soundtrack is my favorite music of all ti...
3	2	Excellent Soundtrack	I truly like this soundtrack and I enjoy video...
4	2	Remember, Pull Your Jaw Off The Floor After He...	If you've played the game, you know how divine...

```
Dataset_1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3600000 entries, 0 to 3599999
Data columns (total 3 columns):
 #   Column    Dtype 
---  ------    ----- 
 0   polarity  int64 
 1   title     object
 2   text      object
dtypes: int64(1), object(2)
memory usage: 82.4+ MB
```

Table of contents

2 Data Cleaning

Duplicated Rows and Null Values

HLT Project: Data Cleaning

- The datasets were already quite clean.
- We had to do only a few operations.

Dropping Duplicated Rows

- We eliminated duplicate rows, keeping one occurrence per row.

```
Duplicated_raws=Dataset_2.duplicated()
print("The total number of duplicate rows are", Duplicated_raws.sum())

Dataset_2=Dataset_2.drop_duplicates()
```

```
The total number of duplicate rows are 931690
```

Deleting Rows with Null Values

- We dropped all rows containing null values in some column.

```
Dataset_2.isna().sum()
```

```
review/score        0
review/summary    188
review/text         2
dtype: int64
```

```
Dataset_2=Dataset_2.dropna()
```

Text Cleaning

HLT Project: Data Cleaning

- Defined a `Clean Text` function:
  1. capable of eliminating special characters in the text.
  2. switching all uppercase letters to lowercase letters.
- Part of our model will be a pretrained transformer that works only with lowercase letters.

Clean Text

```
# Create a function that delete special characters from text and bring it all to lower case
def clean_text(df, field):
    df[field] = df[field].str.replace(r"@"," at ")
    df[field] = df[field].str.replace("_", " ")
    df[field] = df[field].str.replace("-", " ")
    df[field] = df[field].str.replace(r'http\S*', '', regex=True)
    df[field] = df[field].str.replace(r"[^a-zA-Z0-9(),\"'\n\s.!?;:\[\]/%]", "", regex=True)
    df[field] = df[field].str.lower()
    return df
```

- Applied this function to both datasets.

Table of contents

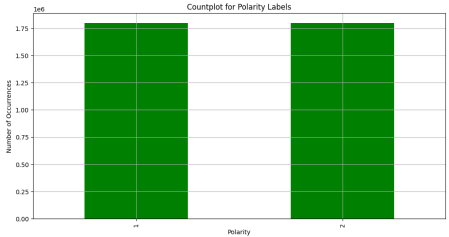
3 Data Analysis

Sentiment's Labels Distribution

HLT Project: Data Analysis

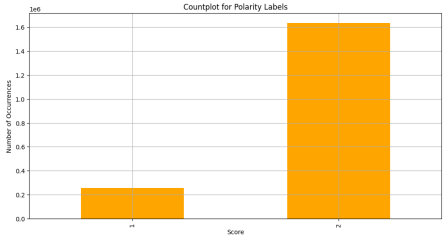
- The first thing we did is to see how the labels were distributed.

Dataset_1



- The labels in the first dataset are already balanced.

Dataset_2



- In the second dataset we have many more positive reviews (about 85% positive).

WordClouds

HLT Project: Data Analysis

- The second analysis concerns the most frequently used words in the datasets.
- Used the library `WordCloud` and defined a function `wordcloud_fun`.

WordCloud

```
# Create WordCloud with text of reviews
def wordcloud_fun(dataset, data_to_plot, figsize=(20,20), max_words=2000,
                  min_font_size=10, height=800, width=1600, background_color="white"):

    plt.figure(figsize=figsize)
    wc = WordCloud(max_words=max_words, min_font_size=min_font_size, height=height, width=width,
                   background_color=background_color).generate(" ".join(dataset[data_to_plot]))
    plt.imshow(wc)
    plt.axis('off')  # Remove the axis
    plt.show()
```

- Searched for the most significant words by dividing texts with negative and those with positive reviews.

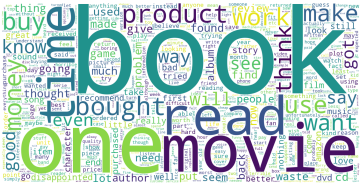
WordClouds

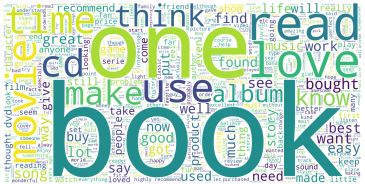
HLT Project: Data Analysis

Polarity=1

Polarity=2

WordClouds on Reviews



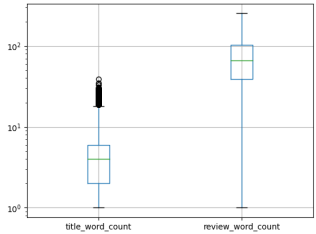


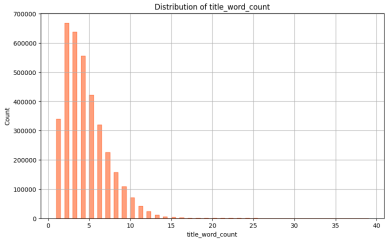
WordClouds on Titles

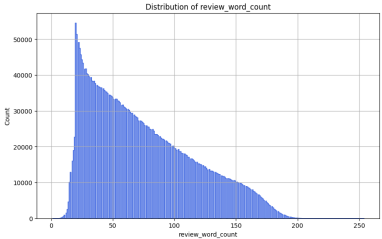
Lenght of Text

HLT Project: Data Analysis

- The last analysis concerns the length of text in reviews and titles.
- Considered the length of a title as the number of words in the title.
- Printed histograms and box plots to study its distribution.







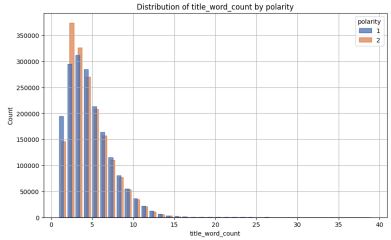
Lenght of Text

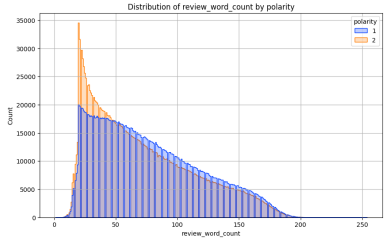
HLT Project: Data Analysis

- At this point we came up with a question:

  "Is there a correlation between the length of reviews (or titles) and whether they are positive or not?"

- Again printed out histograms to try to give us an answer.





- No. The distributions for positive and negative reviews are almost identical.

Bibliography

HLT Project: Progress Update

[1] Angelido. "HLT-Sentiment-Analysis: Sentiment Analysis project repository." *GitHub*. GitHub Repository. (2024)

[2] Wes McKinney. "Pandas: powerful Python data analysis toolkit." *Python for High Performance and Scientific Computing*. *Vol. 14. No. 9*. Pandas Documentation. (2011)

[3] Andreas Mueller. "Word Cloud: A Command Line Interface for Creating Word Clouds." *Journal of Open Source Software 3.26: 781*. Word Cloud Command Line Interface. (2018)

[4] Michael Waskom et al. "Seaborn.histplot: Plot univariate or bivariate histograms to show distributions of datasets." *The Journal of Open Source Software 6.60: 3021.* Seaborn.histplot Documentation. (2021)

HLT Project: Progress Update

Thank you for your attention! :)