HLT Project Proposal

Sentiment Analysis on Amazon Reviews

HLT - Group 11

Angelo Nardone, Riccardo Marcaccio, Matteo Ziboli

Dipartimento
di Informatica
Università di Pisa

Group 11

HLT Project Proposal

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

HLT Project Proposal

Understanding whether a product review is positive or negative is useful for several reasons:

1. **Enhanced User Experience:** Provides valuable information and recommendations about the quality of the product.
2. **Business Insights from Customer Feedback:** Extracts valuable insights from customer feedback, enabling businesses to make informed decisions aimed at improving their products and services.
3. **Effective Online Reputation Management:** Manages their online reputation more efficiently by promptly addressing negative feedback and promoting positive engagement with customers.

Goal of the Project

HLT Project Proposal

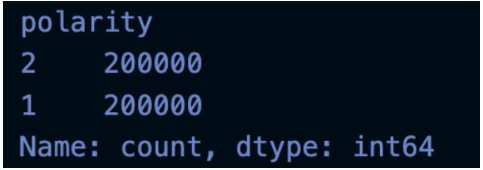
- **Binary Classification of Product Reviews**: Aims to classify Amazon product reviews as positive or negative based on their content.
- **Approaches**: Attempt binary classification using initially only review titles, and then extend our approach to utilize entire reviews instead of titles.
- **Comparison and Evaluation**: Compare the performance of these two classifiers using various metrics to determine the most effective approach.
- **Potential Expansion**: If deemed necessary, explore a secondary dataset to undertake additional classification tasks, using the methodology applied to the first dataset. In this scenario, compare the performance of the classifier between the initial and secondary datasets to assess its effectiveness.

Main Dataset

HLT Project Proposal

- **Available at**: Amazon Reviews Dataset.
- **Context**: The dataset consists of reviews from Amazon. The data span a period of 18 years, including 400.000 reviews. For more information: [1].
- **Content**: The dataset contains 3 different columns: polarity, title, text.

Column	Description
`polarity`	$i = \begin{cases} 1 & \text{if the review is negative} \\ 2 & \text{otherwise} \end{cases}$
`title`	Title of the review.
`text`	Entire text of the review.



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

Figure: The distribution of polarity values in the main dataset.

Second Dataset

HLT Project Proposal

Column	Description
`review/score`	$i = \begin{cases} 1,2 & \text{if the review is negative} \\ 3 & \text{if the review is neutral} \\ 4,5 & \text{if the review is positive} \end{cases}$
`review/summary`	Title of the review.
`review/text`	Entire text of the review.

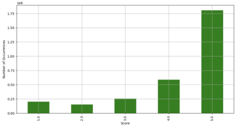


Figure: The distribution of review/score values in the second dataset.

- **Available at**: Amazon Books Reviews Dataset.
- **Context**: The dataset consists of books reviews from Amazon. The dataset contains feedback from approximately 3 million users. For more information: [2], [3].
- **Content**: The dataset contains a lot of columns. We will focus only on three columns: review/score, review/summary, review/text

Implementation and Evaluation

HLT Project Proposal

- **Transformers**: Utilize transformers for binary classification task of sentiment analysis.
- **Lexicons**: Incorporate lexicons as additional features to enrich text representation or as labels for pretraining tasks.
- **Evaluation Metrics**: Assess model performance using accuracy, precision, recall, F1 score, ROC curves and other metrics, comparing also results across different input types.
- **Continuous Improvement**: Plans will evolve as we gain knowledge, ensuring effectiveness and efficiency throughout the project.

Bibliography

HLT Project Proposal

[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.

HLT Project Proposal

Thank you for your attention! :)