CLEVERLY

Introduction

Welcome!

Thank you for your interest in joining Cleverly! We are a small and handpicked team where everyone counts and we created this data science challenge to help you show off your technical skills and have a better understanding of how it would be to work together.

There are no right or wrong answers to this challenge, but we are looking for the following specific skills:

- Problem solving (understanding and structuring the problem)
- Analytical mindset (taking the right conclusions)
- Machine learning and coding proficiency (good domain of ML tools)
- Communications skills (explaining difficult things in an easy way)

Rules & Submission

We value honesty above everything else. Doing this challenge by yourself is the best way for all of us to understand if you are a good fit for the type of work you will be doing.

Our data science team works with Python, but you are free to use any other programming language for the challenge.

The deliverables are the following:

1. Your answers to each of the questions (choose your preferred support)
2. Your code

Finally, we wish you a great time working on this challenge! Do let us know if you have any questions.

Good luck!

Challenge

Most of our work at Cleverly consists in developing natural language understanding models. This challenge consists of analyzing a dataset of online product reviews, which contains a lot of text, a similar domain to the one we work with at Cleverly.

You'll find a list of questions to guide your work in the "Tasks" section.

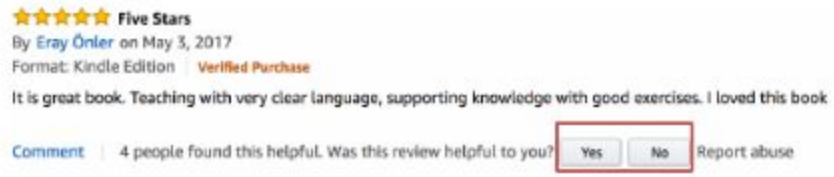
Data

Please download the dataset from this link. It is a sample from the Amazon product reviews dataset, which we downloaded from this page. We extracted a sample from the original dataset so that we minimize the hardware requirements for this challenge.

The dataset is in CSV format and should have the following columns:

- reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
- asin - ID of the product, e.g. 0000013714
- reviewerName - name of the reviewer
- helpful - helpfulness rating of the review, e.g. 2/3
- reviewText - text of the review
- overall - rating of the product
- summary - summary of the review
- unixReviewTime - time of the review (unix time)
- reviewTime - time of the review (raw)

The helpfulness column shows: [total number of "Yes" votes, total number of votes], for the question "Was this review helpful to you?"



Example of product review

Tasks

1. Analysis

We prepared a few introductory questions to get you started with this dataset:

- Is there a correlation between the rating of the product and the helpfulness of the review?
- Who are the most helpful reviewers?
- Have reviews been getting more or less helpful over time?

2. Modelling

Next, we would like to know: after someone writes a review, will it be considered helpful by other users?

Please answer the above question through these tasks:

- Build a model to predict the helpfulness of a review
    - Please approach this as a binary model - you get to create the label yourself.
- How would you evaluate this model?
- Please list some ideas which you would explore to improve the model if you had more time

Note: we won't judge this challenge's result by the performance of your model, but rather by your approach to the problem. The model architectures you use will give us an idea of what kind of architectures you are an expert on.

3. Bonus question

Finally, here are a couple more questions if you are in the mood for extra show off:

- Check out the BERT model architecture. Would you use it to build a new model to predict the helpfulness of a review?
- How do you expect that this new model's performance compare with the previous one (from Exercise 2)? What makes the BERT model better/worse?