Introduction

Welcome!

Thank you for your interest in joining Zendesk! We are a committed team where everyone counts and we created this data science challenge to help you show 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 and understanding of ML concepts and tools)
- Programming proficiency (how you structure and organize your code)
- 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 team works with Python, but you are free to use any other programming language for the challenge.

The deliverables are the following:

1. Project submission
    a. Your answers to the challenge – choose your preferred support
    b. Your code
2. Presentation and discussion of the submitted work with the team
    a. Consider that you will have ~20min to present your approach to the challenge.
    b. Feel free to choose the support that works best for you

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

Good luck!

Challenge

A big part of our work involves developing natural language understanding models. Similarly, this challenge consists of analyzing a dataset of online product reviews, containing a lot of text.

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

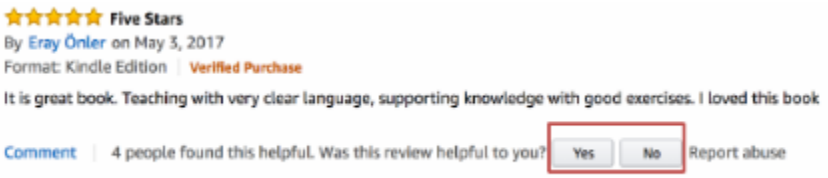
Data

Please download the dataset from this link. It is the Amazon product reviews dataset, which we downloaded from this page.

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

Project

The main goal of this challenge is to answer the main question:

> *After someone writes a review, will it be considered helpful by other users?*

1. Analysis

Data analysis is a relevant stage of a machine learning project, in order to better understand the project and dataset at hand. We prepared a few introductory questions that could help you started with the analysis of 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. Modeling

Please answer the main question of the challenge (*after someone writes a review, will it be considered helpful by other users?)* considering the following steps:

- 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?
- How could you 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. Transformer models

Additional questions:

- Check out the transformer architecture from the BERT model. Would you use it to build a model to predict the helpfulness of a review?
- How do you expect the performance of transformer models to compare with other approaches?