Pattern Product Review Analysis
Statement of Work (SoW) Document

Prepared by	Sehaj Chawla <sehajchawla@g.harvard.edu)>; Taro Spirit <tspirig@g.harvard.edu>; Lotus Xia <lxia@g.harvard.edu>; Heather Liu <yingchen_liu@g.harvard.edu>
Prepared for	Hamilton Noel <hamilton@pattern.com>; Jacob Miller <jacob@pattern.com>; Jed Brunson <jed@pattern.com>

Background

eCommerce is gaining dominance in the global retail market. Due to the lack of in-person interactions on virtual marketplaces, customer reviews become one of the most important channels for customers to communicate with each other and to provide feedback to retailers.

On the one hand, customers rely on reviews from past customers to have a peek at product quality and to make purchasing decisions accordingly. On the other hand, understanding customer reviews is also crucial to the success of online retailers. For instance, retailers can improve product quality or change marketing strategies based on customer preferences and suggestions. Retailers may also rely on review sentiment to predict sales volumes in the immediate future and to determine the quantity of inventories to acquire.

In this project, we worked with Pattern, an eCommerce accelerator, to create a natural language processing (NLP) model that 1) predicts future sales volumes from customer reviews and 2) extracts salient, predictive features from these reviews.

Problem Statement

Our goal is to generate a machine learning model that can predict the success of a product on Amazon (the rank of a product), using data corresponding to the reviews for that product and their related data, i.e. the reviews titles, the reviews ratings, the reviews votes (how many people found that review useful), etc. We will attempt using NLP techniques on the data provided by Pattern, which we will have to pre-process.

For this project, we have two goals: (1) using review data (text and metadata) to predict the future performance of a product, and (2) gaining insights and interpretability for what keywords, topics, or

themes in reviews lead to better or worse future performance.

For task (1), we focus on predicting best seller rank (BSR) as a measure of product's performance relative to other products[1], and this metric is available to us in a time series over a varying number of days for each product. We would like to start by reducing the noise in the BSR time series (example shown in Figure 1), by setting up an averaged out target variable that could represent something like the product's average BSR after n days from the latest review. This sets up the regression problem to solve for our initial model - given the review text and metadata, we will predict the average rank for the future, and to do this, we will make use of transformer models (specifically BERT/GPT2) with additional dense layers to finetune the model for the task at hand. We will then move to a more elaborate model that will incorporate the time series aspect of our target variable by potentially combining the transformer model's dense feature vector with other product data from the past and present under an LSTM structure to predict the future changes in BSR.

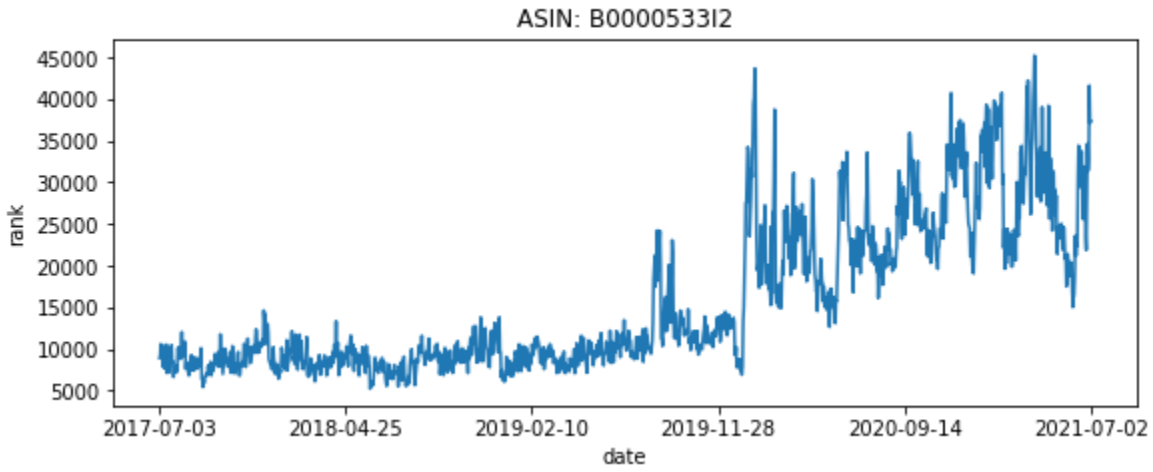

Figure 1: Example of the BSR time series for product with ASIN B000053312

For task (2), we are planning on exploring unsupervised (or semi-supervised) learning techniques (specifically Latent Dirichlet Allocation) for theme extraction from reviews of the product - in order to better understand what the large majority of the reviews for a product are talking about, and what that means for the performance of the product in terms of BSR. Moreover, analysis of the final attention levels given to the review text tokens from our model from part (1) will also be useful in determining predictive strength of specific words in reviews. Lastly, we will also explore supervised learning and transformer models for classification of reviews across distribution shifts, to potentially classify reviews with respect to the complaints customers are making (e.g. delivery, product quality, packaging, servicing, etc.)

Resources

1. Datasets

a. Review data, some information included:
- Review date and name of country where review was left
- Product name and ASIN (product id number)
- Review text and title

[1] Pattern has an approximate model to convert BSR to sales quantity that we might use.

    - Review rating
    - Review votes (how many people found the review useful)
  b. Sales ranking, information included:
    - Date
    - ASIN (product id number)
    - Rank number
    - Average price over the past 180 days

2. Computing Infrastructure

  a. Jupyter Notebook for experiments
  b. AWS for model training

High-Level Project Stages

Our plan is to divide the project into 2 stages, with a focus on the second stage.

1. Using review data (text and metadata) to predict the future performance of a product.

- Processing data and selecting appropriate features for training and testing
- Building transformer models and fine-tuning the model for the task at hand
- Integrating the time-series aspect by potentially combining the transformer model final dense vector with other product data under an LSTM structure to predict the future changes in BSR
- A good prediction model will be conducive to highlight the importance of common themes/words in review text

2. Gaining insights and interpretability for what keywords, topics, or themes in reviews lead to positive or negative future performance.

- Exploring unsupervised (or semi-supervised) learning techniques to understand reviews for a product
- Determining predictive strength of specific words in reviews
- Exploring some supervised learning and transformer model classification on the reviews across distribution shifts, to potentially classify reviews with respect to the complaints they are making

Model Evaluation

We do not yet have any quantitative measures of model performance in mind at the moment, but we hope to achieve the following as we pursue the two aforementioned goals:

1. Determine whether review text is useful in prediction of product performance. In particular, we would like to compare the performance of the model without review texts with the model that incorporates review texts.

2. Obtain a set of sensible words and/or themes from reviews that are predictive of product performance and are consistent with our intuition.

Project Timeline

Sprint ending	Tentative milestone or goal

2022-02-09	• Project set up
	  ○ Git repository created
	  ○ Statement of Work
	• Exploratory Data Analysis
	• Brainstorm project directions

2022-02-13	• Compile list of technical/data and business questions for Pattern • Discuss project directions and select one to pursue
2022-02-21	• Brief literature review • Finish data cleaning • Determine features that would be useful
2022-02-25	• Create a set of baseline models for the first task • Train models on subsets of data
2022-03-01	• Prepare for 15-min presentation • Submit materials on Github • Finish self-/peer- evaluation • Review another team's reports • Finish summary report + technical attachment to partners
2022-03-03	• Milestone 1 Due
2022-03-10	• Task 1: Further improve the model and finetune for the task at hand • Task 2: Create a set of baseline models for the second task
2022-03-17	• Spring Break
2022-03-22	• Prepare for 20-min presentation • Task 1: Get a final deliverable model and train on the complete dataset • Task 2: improve the model and train on subsets of data
2022-03-24	• Milestone 2 Presentation • Submit materials on Github • Finish self-/peer- evaluation • Review another team's reports • Finish summary report + technical attachment to partners
2022-03-31	• Milestone 2 Due
2022-03-30	• Refine models
2022-04-10	• Complete models for both task • Start on blog write-up

2022-04-19	• Prepare for 20-min presentation • Submit materials on Github • Finish self-/peer- evaluation • Review another team's reports • Finish summary report + technical attachment to partners
2022-04-21	• Milestone 3 Due
2022-04-25	• Finish blog write-up • Prepare for final presentation • Initial poster • Impact statement
2022-04-30	• Deliver models and documents to Pattern
2022-05-05	• Final Presentations