How to spot fake review?

Abstract

Using dataset from Amazon, we will analyze how fake reviews affect product ratings and prices. Even though we cannot know which review is manufactured, the verified purchase column from the dataset should give us which review can be considered trustworthy. Comparing the rating, languages between verified reviews and those that are not should help us spot the fake reviews and provide a glimpse of their effect.

Problem Statement and Business Case

Customer review is one of the most important factors for a successful online business. Based on a data from FanandFuel in 2016, more than 90% of customers read reviews before making a purchase. In addition, a research from Northwestern University indicates that people tend to buy a product with reviews 270% more than those without. With the importance of review, it is thus necessary to weed out the fake ones. In this project, we will develop a model to spot out the fake reviews and investigate the impact they have on the products.

Data Science Workflow

- Explore question: What's the structure of the verified review? Most helpful one? How are they compare to the ones that are not verified?
- How do types of product affect reviews? Prices?
- Help customers spot out the fake reviews, establish trust.

Data Collection

amazon_us_reviews | TensorFlow Datasets

This data contains over 130+ million customer reviews on Amazon in TSV file. Each line in the data files corresponds to an individual review. Some helpful dictionaries: rating, verified purchase, helpfulness votes, review description, product category, price.

Data Processing, Machine Learning

- Text normalization
- Explore features that are most important: repeated words, ratings, helpfulness.
- PCA
- Random Forest, Neural Networks

Deliverables

- Jupyter notebook
- 10-12 pages report