

Sentimental Analysis of Amazon Review Dataset

Members:

18075049 Rishabh Bansal

18075052 S Ujjwal

18075062 Vaibhav G

18075066 Yash Bhosale

Problem Statement:

The project contains two parts both on the amazon dataset:

1. Create a model for sentiment classification for the prediction of user ratings based on their reviews using a **self-attention sentence embedding mechanism**.
2. Product Recommendations based on the reviews using **collaborative filtering**.

Description:

The Internet has revolutionized the way we buy products and most of our requirements are fulfilled by the e-commerce world of the online marketplace. In this scenario it is not possible to experience the products before purchase, so most people rely on the product reviews for shaping their opinion before buying the product. Therefore, a model that can summarize the product reviews in terms of the customer ratings can really help in sorting through the items. (**sentiment classification using self-attentive sentence embedding**).

Product recommendation can also be useful for the company as well as the customers as it automatically filters the products based on previous transactions and interests and compels more customers to buy the product thus benefiting both the customer and the company. (**Product recommendation using collaborative filtering**).

Dataset:

The dataset includes electronics product reviews such as ratings, text, helpfulness votes. The dataset is obtained from <http://jmcauley.ucsd.edu/data/amazon/>.

Data Preprocessing:

First, the data needs to be preprocessed, which involves first converting it from JSON to CSV so that it is easy to work on.

The most important part of data preprocessing will involve text preprocessing which inturn involves:

- Removing HTML tags.
- Expanding Contractions (shortened version of words)
- Removing special characters
- Lemmatization
- Removing stop words
- Building a text normalizer

Exploratory Data Analysis:

After performing the preprocessing part we can have a better understanding of our dataset and perform some exploratory data analysis on it. Useful insights can be generated which will further help in creating the sentiment analysis model.

Some basic fields will be:

- Top 10 most reviewed brand and product
- Top 10 least reviewed brand and product
- Most positively reviewed product
 - Its insight words from good reviews as well as insight words from bad reviews
- Most negatively reviewed product
 - Its insight words from good reviews as well as insight words from bad reviews

We will also generate certain analysis from the e-commerce company point of view such as

- The year with the highest number of customers
- The year with the highest number of reviews
- Average helpfulness of products
- Distribution of review length
- Distribution of helpfulness and review length
- Top words for good reviews
- Top words for bad reviews

We will get useful information as to which is the best-reviewed overall product in a category as well as the worst-reviewed. This can help customers get an overview as to which product is the best/worst in its respective categories.

Also, companies will also get insight as to which product is more popular among the masses and which on which category they may consider giving a discount.

Sentiment Classification using Self-Attentive Sentence Embedding:

This model consists of two parts:

1. A Bidirectional LSTM
2. The self-attention mechanism for sentence embedding

The self-attention mechanism provides a set of summation weight vectors for LSTM hidden states. The set of summation weight vectors are dotted with the LSTM hidden states, and the resulting weighted LSTM hidden states are considered as an embedding for the sentence.

Now, this embedding can be combined with the multilayer perceptron terminating with a softmax layer for sentiment analysis.

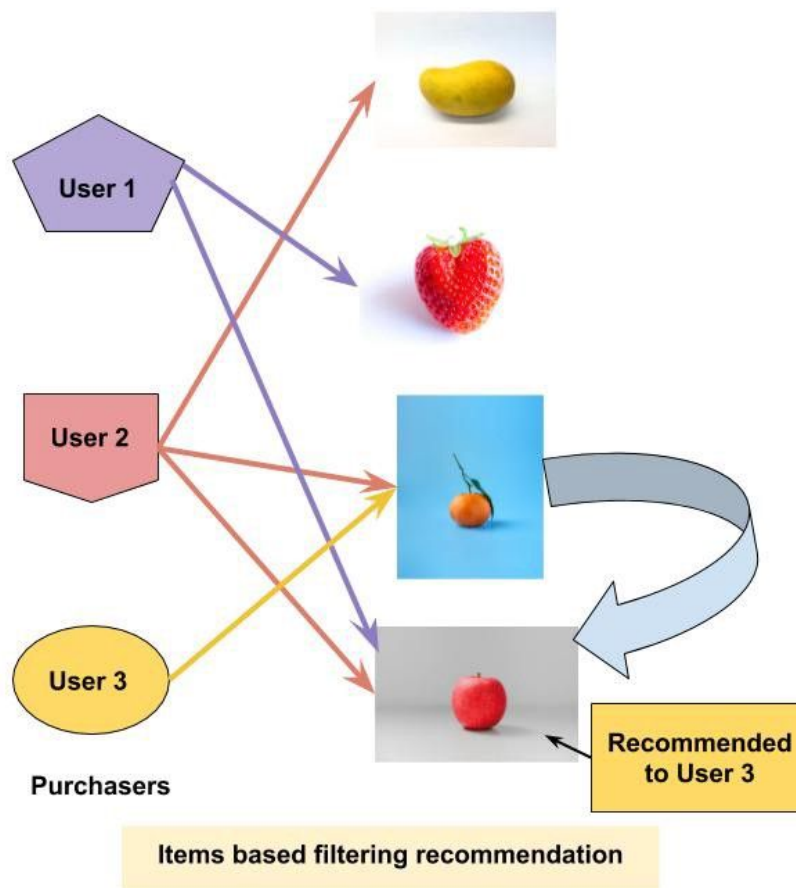
Product Recommendation based on Collaborative Filtering

People usually buy products based on recommendations from their peers and friends. Also, a major factor in the decision-making process is the rating of a product online. We see that there are similar trends in buying patterns among like-minded people.

We will design a recommendation system that will first capture the past behavior of a customer and based on that recommend products which he is more likely to buy. We intend to use item-item collaborative filtering as the number of users is definitely more than the number of products available. We use the principle that if users have bought similar products in the past then they will probably agree in the future.

In the filtering, we will calculate the similarity between each item pair, and based on that we will suggest similar items that were liked by users in the past.

A pictorial representation is shown below:



If we can recommend a few items that the customer likes, it can greatly improve the user experience for our e-commerce platform.