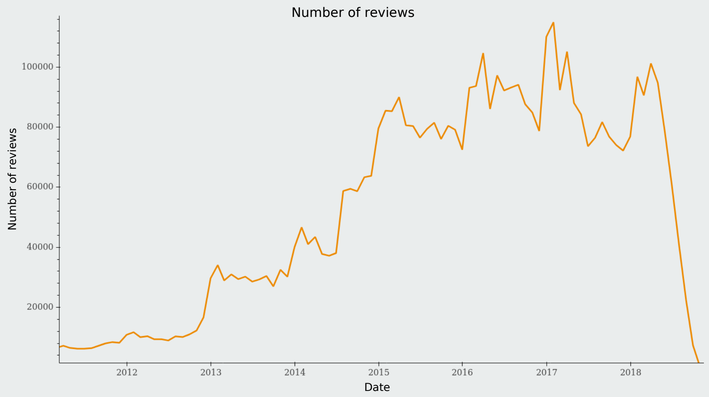
Impact of events and trends on Amazon

Jules Gottraux, Lucien Iseli, Florian Ravasi and Christina Mantonanaki

The dataset contains information about products from Amazon's *Grocery & Gourmet Food* category along with their reviews. Our findings are mainly, but not only, based on the text of the reviews.



A few statistics:

- 5 million reviews
- 2.7 million reviewers
- 5GB of text and metadata
- More than 90% of reviews between 2013 and end of 2018
- Data collection stopped throughout 2018

Here is how the pipeline works:

Amazon Customer

yummy brownies and cookies

May 30, 2018

Size: 18 Pieces | Verified Purchase

These brownies and cookies are very tasty. I would order them again. They also remained fresh for several days.

One person found this helpful

Helpful | Comment | Report abuse | Permalink

1. Lower-case
2. Remove numbers and punctuation
3. Tokenization
4. Lemmatization
5. Remove stop-words

Amazon Customer

yummy brownies and cookies

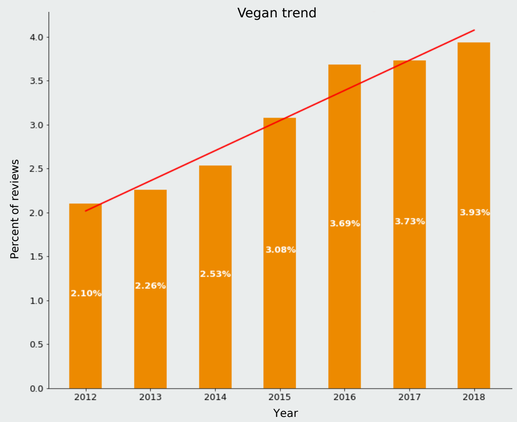
May 30, 2018

Size: 18 Pieces | Verified Purchase

brownie cooky tasty would order also remain fresh several day

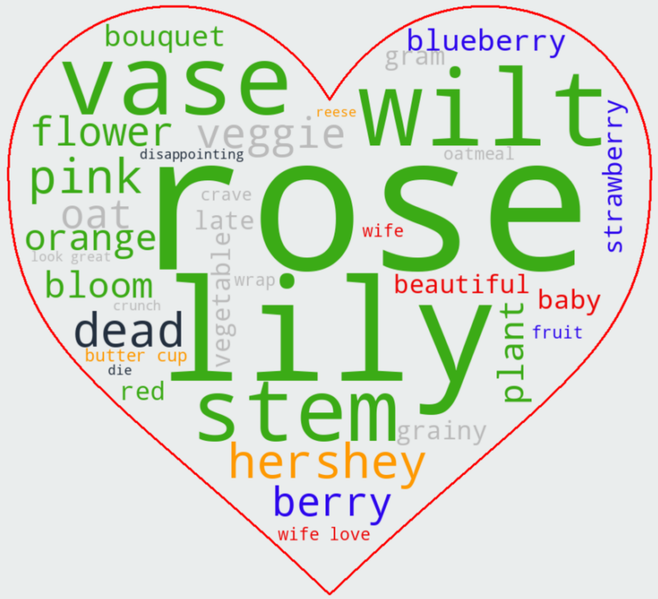
One person found this helpful

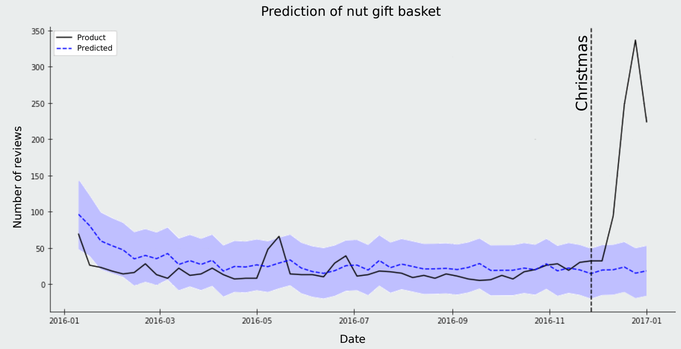
Helpful | Comment | Report abuse | Permalink



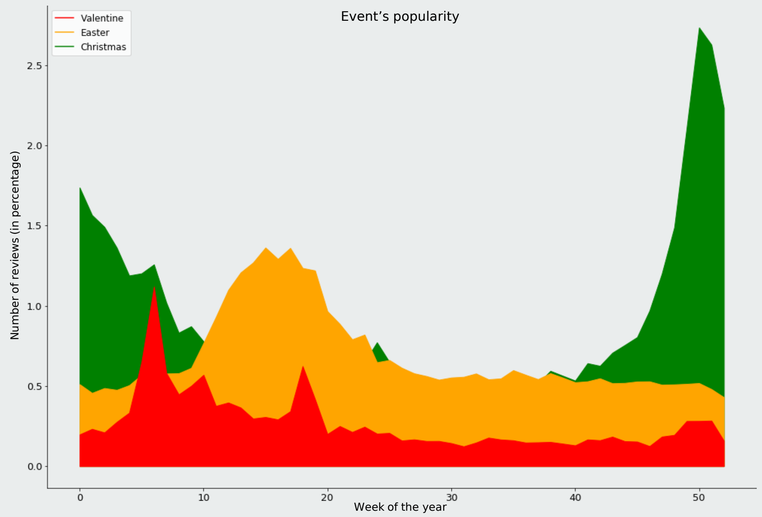
The first step of our analysis consists in identifying trends that evolved over time. Here, we take a look at veganism and notice an increase from the start of 2012 until the end of 2018. Additionally, with a linear regression we test whether the slope is null and it confirms our intuition since it yields a p-value close to 0.

A second axis of our analysis is to measure the impact of events in our dataset and draw conclusion from that. Here is the case of Valentine's day. We measure how much the words characterize the event using the evolution of their frequency. We plot the most important words for Valentine's day in a word cloud, each color represents a theme. We can see how these themes relate to Valentine's day.





Whereas the previous analysis shows how events or trends affect the user's reviews, we can also ask ourselves how products are affected. The goal is to see how Christmas affects products' number of reviews. In order to do so, we use structural bayesian time series. We choose one product we want to study and other related products that shouldn't be impacted by Christmas. We decide to look a Christmas nut-gift basket and compare it with other nuts or chocolate-nut bars. Finally the blue area around the predicted curve symbolizes how sure the model is about the prediction. Therefore, we can notice that without the impact of Christmas, the product we chose would not have met such a high popularity in December.



During this project, we analyzed three events, namely Valentine's Day, Christmas and Easter. From the words that characterize the events, we can measure how much the event is present on Amazon's *Grocery & Gourmet Food* at a given time. This gives us a view on the popularity of these events throughout the year. We can see how Christmas dominates the two other events.