Was the healthy lifestyle trend predictable?

Dario Anongba Varela dario.anongba
Valentin Moullet valentin.moullet
Alain Birchmeier alain.birchmeier
@epfl.ch

Abstract

A recurrent subject has been on the front page of the news on social media, newspapers and in general on the Internet: the subject of healthy and responsible lifestyle. From personalities encouraging others to adopt a vegan lifestyle to fitness applications and YouTube channels, there is a real enthusiasm around it [1, 2]. Our report shows that, using only Amazon reviews data from 2003 to June 2014 [3], it would have been possible to predict the growth of this trend with a high confidence. By applying a similar analysis with more recent data and maybe with other types of products, we expect that it would be possible to discover other possible trends and different insights that are useful, especially for selling or reselling companies.

1 Introduction

> *"A healthy lifestyle, excluding any damaging influences, defines the positive and voluntary measures a person can implement to maintain good mental and physical health. This includes healthy habits in terms of diet, treatment of the body, sex, and the environment"* [4]

The idea of the project is to determine if it was possible to predict the ascending enthusiasm over healthy products with data from 2003 to June 2014.

In order to achieve that goal, we will use the Amazon reviews dataset. From those reviews, we can extract information such as consumer satisfaction, dates of reviews, product categories and the enthusiasm for certain products.

With the emergence of new institutions promoting healthy lives (like vegan shops or fitness centers) or personalities getting more involved in promoting healthy behaviors, finding insights and patterns in people's shopping behavior could be useful to define in which direction this social change is heading, and if selling companies could have predicted this to adapt themselves in order to increase their revenues by, for example, changing their products or adapting their prices and advertisement.

2 Related work

While looking for other related work, we realized that it was hard to find a study similar to ours. Most experts focus on what will be the next healthy trends (e.g. yoga, healthy spices, healthy pet foods, ...), like [5]. Our study is unique in this sense: we are trying to find if the healthy lifestyle trend was possible to be predicted early, in 2014, by looking at what products people were reviewing on Amazon. We've found no other significant data analysis research on this.

3 Data Collection

3.1 Dataset description

The data retrieved from Amazon is in JSON format. It contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 to July 2014.

This dataset includes reviews (reviewer id, product id, ratings, review text, helpfulness votes and review date), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).

For more information about the original dataset, see reference [3].

3.2 Data pre-processing

The first task was to collect the Amazon data, organize it and store it in a ready-to-use data format (pickles).

Extracting healthy products

Since this analysis is about healthy lifestyle, we mostly focused on products and reviews from two main Amazon categories: **Grocery and Gourmet Food** and **Sports and Outdoors**. In order to extract products that can be considered healthy from those categories, we used two different methods.

The first method, applied to the Grocery and Gourmet Food category, was to use a keyword-based approach. By scanning the product description and the product title, we filtered products that did not contain at least one of the keywords. This filtering was necessary for that category because it could contain a mix of products that are unhealthy or that don't have any particular reason to be considered healthy. The used keywords were:

- ”organic”, ”natural”, ”sugar-free”, ”healthy”, ”vitamin”, ”supplement”, ”minerals”, ”diet” and ”vegan”

The second method, applied to the Sports and Outdoors category, was to select entire sub-categories that represent a healthy lifestyle. In this case, we do not need to filter particular products, but only categories. The chosen categories were:

- ”Exercise & Fitness”, ”Cycling”, ”Sport Watches”, ”Team Sports”, ”Strength Training Equipment”, ”Action Sports”, ”Cardio Training” and ”Running”

Finally, given those filtered products, we extracted all related reviews. By this method we obtained a total of 1’332’644 healthy product reviews, from which 467’338 are from the Grocery and Gourmet Food category and 865’306 are from the Sports and Outdoors category.

Extracting active reviewers and products

An important assumption made in this analysis is that users that are enthusiasts over certain products tend to comment them (or simply give a rating) more than users that are not particularly involved in that type of products. So, if a type of product is trendy, we should see the number of involved people increase, resulting in an increase in the number of reviews.

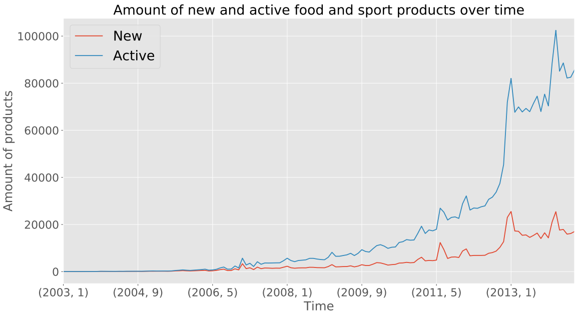

Figure 1: Amount of new and active Food and Sports products over time

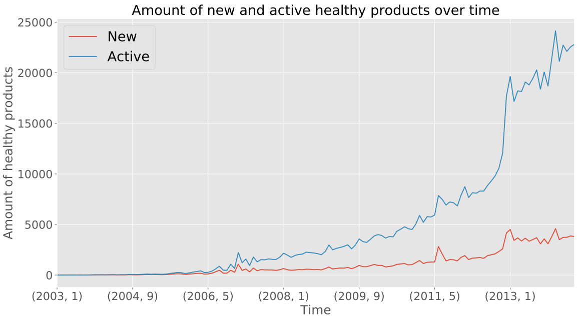

Figure 2: Amount of new and active healthy products over time

In order to determine if that assumption is true for healthy products, we need to extract the active products (products that have at least one review) over time. We use the tuple (year, month) as the unit of time. Extracting this data is necessary because it would be wrong to assume that a product that has been active at a given time, will still be active in the future. For example, a product having been reviewed in June 2004 shouldn’t be taken into account as an active product in August 2007 unless he has also been reviewed that month. This is the only way to determine, given our data, if a product is currently active.

Figure 1 shows the number of active products per month for the Food and Sports categories and Figure 2 shows the number of active products per month for the healthy products.

Note that we can see a huge increase of reviews at the end of 2012. We will see this pattern in almost all our graphs, and after doing some research, we saw that it correlates with the launch of Amazon in Brazil which is a huge market opening [7].

3.3 Summary statistics

Diving into the data, we performed a descriptive analysis in order to extract useful information.

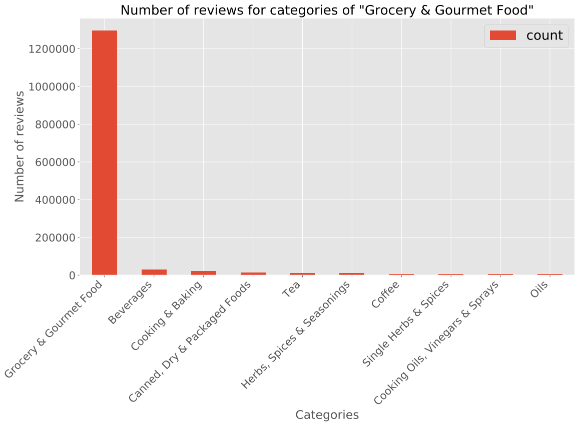

Figure 3: Amount of reviews in the Grocery and Gourmet Food sub-categories

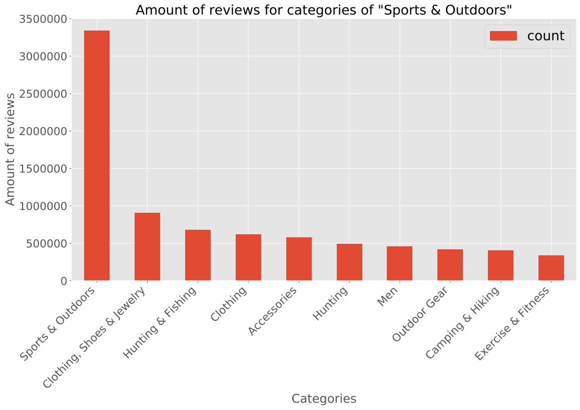

Figure 4: Amount of reviews in the Sports and Outdoors sub-categories

Number of reviews per category

We can see in fig 3 the number of reviews per sub-category of the main Grocery and Gourmet Food. The figure shows that most of the products in that category are not assigned a specific sub-category. Knowing this helped us avoid using whole categories in the filtering process.

Similarly, we can see in fig 4 the number of reviews per sub-category of the Sports and Outdoors root category are well balanced and that it is possible to use entire categories in the filtering process.

Restricting the time range

While doing the descriptive analysis, we realized that we needed to restrict the time range because most categories didn't exist or didn't have enough products. We decided to use reviews from 2003 to June 2014.

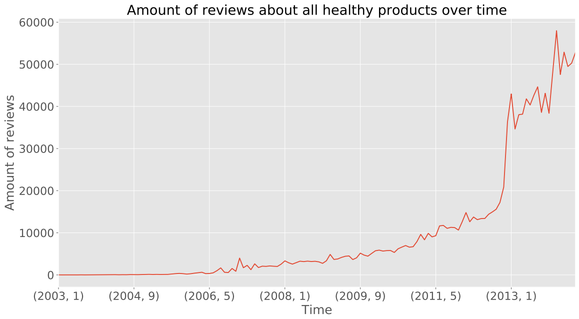

Figure 5: Amount of reviews about healthy products over time

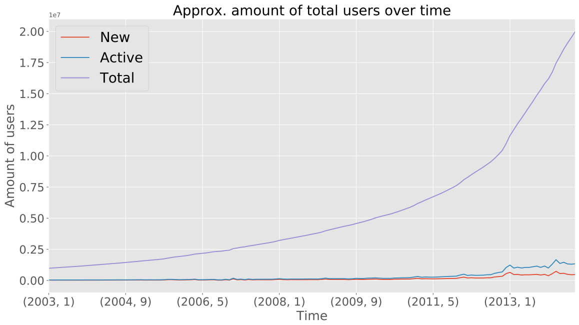

Figure 6: Amount of users seen on Amazon

Reviews over time

As an analysis, we could have shown that the number of reviews about healthy products grows exponentially over time (see Figure 5). From this, a naive interpretation would be that the enthusiasm for healthy lifestyle has grown exponentially. Obviously, this interpretation is wrong because it doesn't take into account the potential exponential growth of Amazon, as Figure 1 and 2 suggest. To prove it, we decided to compute the total number of users seen on Amazon. The results are confirming this growth as you can see in Figure 6.

We will need to take into consideration the global growth of Amazon over time (number of products, number of reviewers and number of reviews) in comparison to the same metrics for the healthy products only, in order to conclude if yes or not, the healthy lifestyle trend was something predictable.

4 Core concepts

The purpose of the project is to determine whether or not the ”hype” for healthy products is growing over the years. In order to achieve that goal, **we define the ”hype” as the average number of**

reviews per active product for a certain type of products. For example, we would say that DVDs are more ”hyped” in 2013 than Blu-ray if we see that the average number of reviews for DVDs is in general bigger than the average number of reviews for Blu-ray in 2013.

We have to be very careful with this definition: the way we are defining hype works for comparing products during the same time period, but you couldn’t say that DVDs are more hyped in 2013 than in 2004 even if the average number of reviews for DVDs in 2013 is bigger than in 2004, because it is also dependent on other factors (number of users, policy for reviewing at this time, etc).

Considering the different nature of products on Amazon, it wouldn’t make sense to compare the hype for healthy products against the hype of the whole Amazon. Indeed, some products tend to be naturally more reviewed than others, like books or movies. Given that, in our analysis we only compare healthy products with their corresponding entire categories.

5 Results and findings

In order to answer our initial question, we want to compare the hype of healthy products with the hype of all products in the Grocery and Gourmet Food and Sports and Outdoors categories.

We want the ”hype” for each month from 2003 to June 2014. To do so, we first need to compute the number of reviews in the Grocery and Gourmet Food and Sports and Outdoors categories for each month. This can be seen in figure 7.

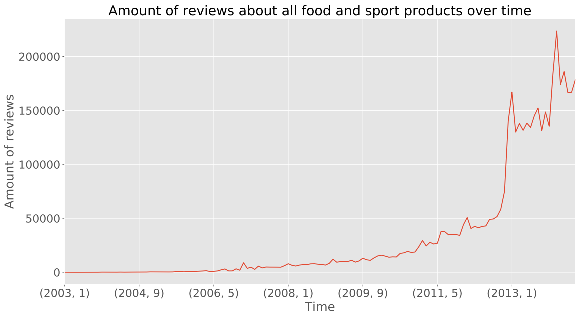

Figure 7: Amount of reviews of Food ans Sports products

Now that we have all the information we can get the hype comparison seen in figure 8.

Please recall the warning about our hype definition: we cannot say that products are more hyped after 2013 than before just because the average number of reviews per product is higher, the number of reviews is dependent on other factors.

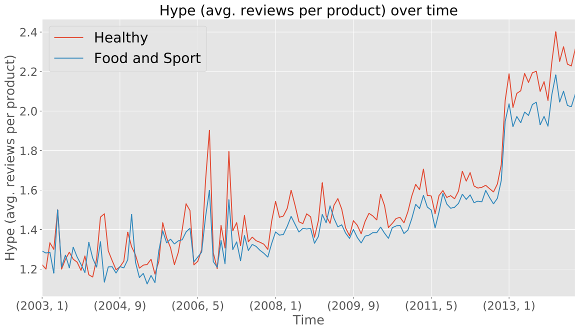

Figure 8: Hype (avg. reviews per product) over time

But still, this plot is quite interesting. It seems that before 2006-2007, it’s hard to tell which type of product is more hyped, but since the end of 2008, we clearly see that in general the healthy products are more hyped than the others.

It is also interesting to see the ratio between those values for each month in order to see the evolution over time. It is important to compute the ratio and not the difference here, because as the average number of reviews goes higher with time, the difference would also grow by default. For example, if the values are 1 and 0.5 in 2003 for healthy and other products respectively, and 3 and 2 in 2013, it would be wrong to say that the ”difference” of hype for healthy products and general products is more visible in 2013 than in 2003, even though the actual difference between the numbers is bigger in 2013; we should compute the ratio instead, and we would see that the ratio in 2003 would be bigger than in 2013.

Finally, we show this ratio in figure 9 along with a regression on the points that it gives us for every month to see how the tendency evolves. Note: the regression is computed using **Ridge regression, with polynomial degree of 4.** [6]

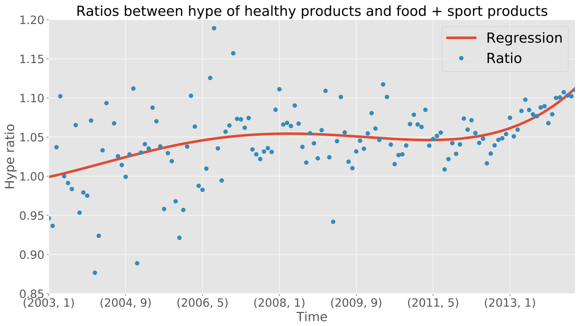

Figure 9: Ratio between hype of healthy products and food + sport products

This graph shows how much hyped are healthy products compared to general food and sport products, and how this tendency has evolved from 2003 to June 2014. We clearly see that the tendency goes higher after 2012, which is indeed what one could have expected with everything we now know in 2017. We can also note that the variance of this ratio is getting lower with time, and this is due to the growth of Amazon and the number of reviews, the more reviews and active users/products there are, the less variance there will be.

6 Conclusion

In conclusion, our analysis shows that it would have been possible to predict the trend of healthy lifestyle with the Amazon reviews data from 2003 to June 2014. Indeed, we saw that the hype for those products was getting drastically bigger compared to other standard products starting from late 2011, and there was no reason for it to stop growing. There is one caveat though: we showed that the hype for healthy products was getting bigger, but we defined some products (roughly one third of them) as healthy based on the fact that they contain some supposedly healthy keywords. We think this is a fair way of categorizing them, but it's possible that people writing descriptions and titles of products were getting more eager to write such words in order to make users buy them more. A more in-depth and precise analysis could be done to categorize healthy products better, but it would be hard to be sure that the descriptions and titles written by people are not biased by the fact that they want people to click on the product and buy it.

References

[1] Nancy Gagliardi, Forbes Contributor
2015 (accessed DEC 11, 2017)
Consumers Want Healthy Food.
https://www.forbes.com/sites/nancygagliardi/2015/02/18/consumers-want-healthy-foods-and-will-pay-more-for-them/.

[2] Deborah Weinswig, Forbes Contributor
2017 (accessed DEC 11, 2017)
Wellness Is The New Luxury.
https://www.forbes.com/sites/deborahweinswig/2017/06/30/wellness-is-the-new-luxury-is-healthy-and-happy-the-future-of-retail/.

[3] Julian McAuley, UCSD.
2016 (accessed OCT 28, 2017).
Amazon Product Data.
http://jmcauley.ucsd.edu/data/amazon.

[4] CCM, Groupe Figaro.
2017 (accessed DEC 03, 2017).
Healthy Lifestyle: Definition.
http://health.ccm.net/faq/3193-healthy-lifestyle-definition.

[5] Sophie Miura, MyDomaine
2017 (accessed DEC 11, 2017)
The 8 Major Health and Wellness Trends of 2017.
http://www.mydomaine.com/wellness-trends-2018.

[6] ARTHUR E. HOERL AND ROBERT W. KENNARD.
1970.
Ridge Regression: Biased Estimation for Nonorthogonal Problems.
TECHNOMETRICS

[7] Zack Whittaker, ZDNet
DEC 2012 (accessed DEC 18, 2017)
Amazon launches in Brazil.
http://www.zdnet.com/article/amazon-launches-in-brazil-opens-kindle-store-tablet-coming-soon/.