Amazon: 20 years of reviews

Lucas Zweili lucas.zweili@epfl.ch
Przemyslaw Juda przemyslaw.juda@unine.ch
Gianni Giusto gianni.giusto@epfl.ch

Abstract

The project focuses on the Amazon's dataset and aims to infer whether reviews that make Amazon's success are reliable or not. To do so, several machine learning algorithms were implemented as well as text mining techniques to determine underlying features that make a review bad or good. The final model trained on methodically chosen features performed well at discriminating reviews qualitatively different from the average opinion.

1 Introduction

Amazon was founded in 1994 and started to sell books. It then diversified to sell various kind of items as we know it today. Since its foundation, the company has collected millions of reviews for each article available on their website.

These reviews are collected daily from people using Amazon's services and give appreciation about someone feeling regarding an item bought on the website. They played an important role in the development of Amazon. Why so? Because they are the key in the purchasing process. Indeed, one often feel more comfortable and safe when buying a product which is well rated? Therefore, it is not surprising to find fake comments as they can potentially impact economical outcomes.

This project aims to explore the reviews across time and distinguish between reliable and fake ones.

2 Related work

As language analysis becomes more and more popular to extract meaningful information from text, several studies were carried on the subject. Most of them used sentiment analysis techniques such as n-grams and bag of words (Fornaciari, Poesio, 2014) where language processing is often combined with machine learning techniques like support vector machine (SVM), K-nearest neighbors etc. Deep convolutional network also came into play by performing sentence classification task for example (Le Cun, Conneau and Schwenk, 2016).

3 Data collection

The dataset is provided by Amazon.com[1] and contains over 130 millions of reviews written on the online platform in the period from 1994 until 2015. The original set contains a collection of US reviews and also a smaller collection of multilingual reviews. In our project we dealt only with the US reviews set, which is composed of reviews grouped by 43 categories, such as: Books, Wireless, PC, Toys, Shoes, Luggage, etc. Each category contains a collection of review records and each record is composed of metadata (the id of review author, the product id, number of positive and total votes about the review from other customers) and the review itself (including the headline and star rating of the product on a scale from 1 to 5).

4 Data description

The raw data were originally in *tsv* format: tab separated value, a text format. Given the size of the dataset (over 30 GB), we performed the initial filtering of the data using `spark` and extracted the features of interest, such as `review id`, `product category` and `helpful votes`. We left out the review texts, not required for extracting the statistics about the temporal evolution of the reviews by category. It enabled us to use `pandas` to do this basic analysis, as we reduced the dataset to around 2 GB. Parquet format was used to store the reduced dataset[2].

[1] https://s3.amazonaws.com/amazon-reviews-pds/readme.html

[2] https://drive.google.com/open?id=1CkC4OMzkUiwjt7un_K6jhMlBVD1L4t5o

Reviews content of the *PC* and *Wireless* categories were used to perform language analysis[3]. The data frames were matched using the review ID as key.

5 Methods

5.1 Amazon temporal evolution

From the data stored in parquet file, `pandas` was used to extract meaningful information from Amazon's expansion. From features listed in section 4, different metric such as mean rating, variance in the ratings or mean reviews usefulness were computed to allow comparison between categories.

5.2 Reviews statistical analysis

First analysis were based on the ratings to see whether there were differences or similarities in the stars' distributions between helpful and not helpful reviews.

The idea here is to perform an observational study. Reviews are assumed to be of comparable quality regardless of the given score.

Data was filtered to keep only reviews that received at least 5 votes qualifying them as useful or not in order to reduce subjectivity in the appreciations.

5.3 Text mining

This part focuses on the reviews' contents and aims to determine what makes a comment reliable or not by building a classifier trained on selected features.

5.3.1 Features selection

Before performing further analysis, selection of a subset of reviews was done.

Criterion selection was again based on the number of votes gather by the reviews which had to be larger than 5.

First, the *usefulness* of the review was taken into account. Indeed, not useful reviews are more prone to be fake because non-constructive or different from the average judgment.

Secondly, the *deviation from the average rating* and the *average deviation per reviewer* seems to be appropriate to quantify the divergence between opinions, as fake reviews tend to attribute notes different from the mean to introduce bias.

[3] `https://drive.google.com/open?id=1yzH01tOqDSTRQm_s1Ex0ATbeybDkzbUK`

Reviews' *length* was also used as feature since fake reviews could have less pertinent arguments as they are not inspired by true facts.

It was also assumed that bad reviews aim to introduce bias in the general opinion. Therefore, they are more susceptible to give extreme positive (5 ★) or negative (1 ★) ratings.

Besides star rating other reviewer-based features were analyzed: the total number of reviews, the average rating, the review's length frequency and the average review's length.

An additional metric based on the text content was calculated using sentiment approaches (c.f. section 5.3.2).

5.3.2 Sentiment analysis

In depth text analysis using sentiment-based approaches allowed us to quantify the information content of a review. Is someone's point of view more positive or negative?

To do so, two approaches were implemented. First, AFINN model was used. The latter attributes a score between -5 and 5 to each English words.

Secondly, sentiment carried by a text-based comment was computed using the Vader module of the `nltk` Python's library. The latter attributes a score of positiveness, negativeness and an aggregate of both, namely *compound score*. This metric sums lexicon ratings and normalize them between -1 (extreme negative) and 1 (extreme positive).

Therefore, for both techniques, the key lies in the overall sentiment of the comment: if the reviewer uses an accentuated positive or negative vocabulary, the review can be spotted easily.

Reviews from *Wireless*, *PC* and *Automotive* categories were subjected to sentiment analysis due to their appropriate size. Indeed, larger categories such as *Books* are too demanding in computational time. Entire data and a subset containing bad reviews were used for comparison. These latter were selected using previously mentioned features, namely: deviation from average rating, helpfulness and extreme ratings.

5.4 Model and classifier

As no set containing fake reviews was available, unsupervised learning methods such as *K-means* and *DBSCAN* were used with the Python's library `scikit-learn`. The classifier was build and trained on a data set of 3 fused categories (*PC*,

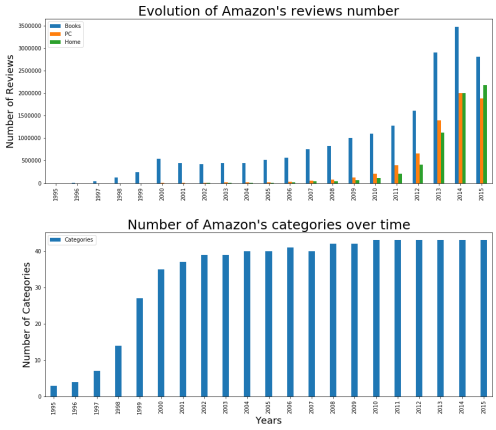

Figure 1: Evolution of the number of reviews for the categories *Books*, *PC* and *Home* and the evolution of the number of categories on a 20 years time window.

Figure 2: Rating distributions comparison between all reviews of the *Books* category and reviews qualified as not useful.

Wireless and *Automotive*) with the features mentioned in section 5.3.1.

6 Results and findings

6.1 Amazon temporal evolution

As seen in Fig.1, Amazon rapidly diversified and proposed items different from books. While the number of categories rapidly increased, the number of reviews remained relatively constant in the beginning.

Only since the last past years, this amount hugely increased following an exponential growth. This progression emphasizes the need of carrying further review-based analysis.

Categories exploration helped us in our attempt to find data to focus on (i.e. more susceptible to contain fake reviews). For example, electronic categories display highest variance in the ratings and are thus more controversial.

6.2 Reviews statistical analysis

In Fig.2, it appears that there are much more negative ratings in reviews which seem not to be useful compared to all reviews where 5 star ratings are most common. From this first insight of reviews analysis, two conclusions can be made: either people tend to be less effective in communicating negative opinions, either bad reviewers give deliberately bad grades to bias the item's mean rating. In the latter case, further analysis must be carried on the reviews' content to determine if they are truly fake.

6.3 Text mining

From Fig.3, one can notice that the majority of reviews achieved high positive score (toward 1) whereas bad reviews sentiment distribution looks different with a higher proportion of negative score. It emphasizes the fundamental content difference of potentially fake comments which tend to be more negative and suggests that the feature *sentiment score* can be relevant in building a classifier. Note that the original dataset was sampled to reduce computational time.

6.4 Final model

After training K-means model, it performed well at discriminating reviews mostly by theirs length although it was not our prior goal, our priority being the separation of bad reviews from the others. Thus, a DBSCAN model was trained following the curse of dimensionality, namely that the distance between points increases with the number of dimension. As DBSCAN try to define regions with "high density", reducing the dimension with principal component analysis (PCA) to improve the performance of DBSCAN seems a good way to do. Note that only features accounting for 90% of the total explained variance were kept.

The optimization of the hyper-parameters $epsilon$ and the $min_{samples}$ which consist respectively in the maximum distance between two sam-

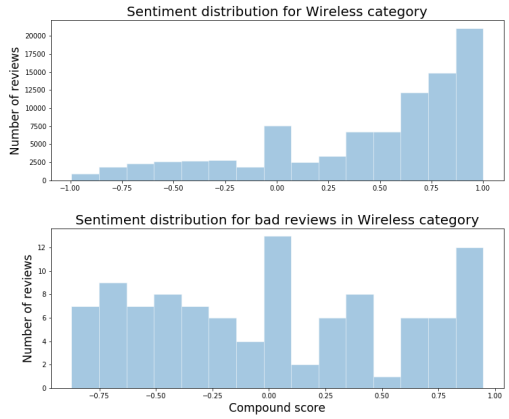

Figure 3: Compound score distribution of all reviews of the *Wireless* category compared to bad reviews from this same category.

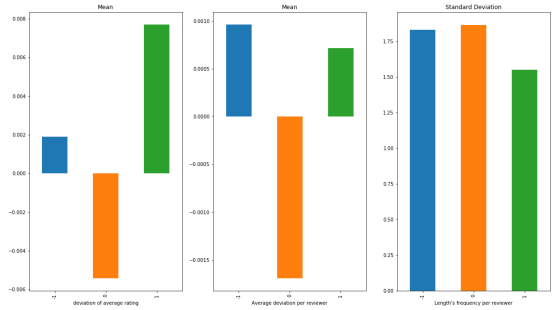

Figure 4: Relevant differences between the 3 main clusters identified by the *DBSCAN* model.

ples and the number of samples in a neighborhood for a point to be considered as a core point were tuned 'manually'. This process ended up with $epsilon = 0.5$ and $min_{samples} = 50$.

The cluster including fake reviews is expected to have a higher standard deviation regarding the *length's frequency per reviewer* as fake reviews are often duplicated. However, a bigger *average deviation per reviewer* can occur as bad reviewers try to induce bias. Therefore, they would have a bigger *deviation of average rating*.

Fig.4 shows the relevant differences between the 3 main clusters identified by the model. Cluster "0" seems to include the majority of the fake reviews considering that it includes all the previously mentioned specifications. However the results should be taken with care. Indeed, due to the wide variety of possible comments and possible way to write fake reviews but also because of subjective ratings, some false negative can appear.

One spotted fake review:

> *This unit is GREAT for laptop and also Desktops! With the One Touch OS backup/files makes it easy with the One Touch button and software provided! [...] You can also boot from Hard Drive for recovering the whole drive! Supports up to 2TB.*"

The author of this comment has graded each of his reviews with a rating of 5. Link to the accentuated positiveness of the word employed, it is highly probable that the this advertising tries to influence the product's rating.

Still, the results given by our classifier were not convincing enough due to lack of accuracy although some fake reviews could be detected. We thus thought about implementing a semi-supervised model but it turns out to be very demanding in time as reviews must be labelled.

7 Conclusion

In this investigation, several complementary methods such as machine learning-based model and text mining were combined to figure out the characteristics that make a review bad. Data was handled successfully and meaningful information was extracted from it.

Our classifier turned out to be efficient in spotting quickly written reviews which intend to introduce bias in the general opinion. However, it remains a challenge to discriminate genuinely formulated reviews.

The literature give very little information about the estimate proportion of fake reviews on Amazon. It is therefore difficult to test the performance of our model.

References

Le Mans University. Conneau A. Schwenk H. Le Cun Y. 2016. *Very Deep Convolutional Networks for Natural Language Processing*.

Conference of the European Chapter of the Association for Computational Linguistics. Fornaciari T. Poesio M. 2014. *Identifying fake Amazon reviews as learning from crowd*. 30(4):279–287.

Ecole de Technologie Suprieure Montral. Elmurngi E. Gherbi A. 2017. *Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Techniques*.