Ebooks vs Books: Comparative User Experience Time Analysis

Etienne BONVIN **Damian DUDZICZ** **Xavier PANTET**

{etienne.bonvin, damian.dudzicz, xavier.pantet}@epfl.ch

Abstract

This report presents the results obtained during the ADA project of 2018 which focused on the analysis of Amazon customers behavior regarding reading. With the help of Amazon books and Kindle ebooks reviews dataset spanning from 2008 to 2014, we explored the impact of the new reading medium on consumer habits. We compared the overall experience of the users for each support. We look for the presence of a possible preference for one or the other and its evolution through time with the help of natural language processing techniques.

1 Introduction

The Amazon Kindle is one of the most popular e-reader on the market, since its launch in October 2007. Various study were made regarding the benefits or the disadvantages of the product regarding the efficiency and quality of the reading on such devices e.g (Merga, 2015). However the majority of such studies involves surveys on small samples of specific population groups and at a given time. Our approach consists of using the extensiveness of the Amazon datasets that span from 2008 to 2014 to observe the impact of the Kindle on users' reading experience and habits.

2 Prior Work

A previous year group had a similar project to the one presented in this report (Lee, Jolles, Lamonato, 2017) but their analysis was made on corresponding books and ebooks. The authors explore different methods in order to match the two formats of a given publication. Unfortunately, the process appears to be very tedious and complicated due to Amazon detection of web scrapping and the limitations of the AWS API's product referencing methods. We decide therefore to adopt a different method based on comparison on the users overall books and ebooks consumption instead on a given product.

3 Used Datasets

3.1 Datasets Collection

We use four Amazon reviews datasets made available on Julian McAuley's (UCSD) webpage[1]. These consist of:

1. `reviews_Books.json` (20GB)
2. `reviews_Kindle_Store.json` (2.2GB)
3. `ratings_Books.csv` (900MB)
4. `ratings_Kindle_Store.csv` (130MB)

3.2 Datasets Description

The two first datasets contain the reviews corresponding to the given medium. They contain extensive 9-tuples consisting, among others, of the review text along with the `reviewerID`, the `ASIN` Amazon product identification number, the `reviewTime` in the UNIX epoch format and the `rating` of the product in the 1 to 5 stars scale.

The last two are simplified versions of the previous corresponding datasets. They only present numerical 4-tuples constituted of the `reviewerID`, the `ASIN`, the UNIX `reviewTime` and the `rating`.

The key aspect of the datasets that makes the analysis possible is the consistency of the `revierID` among all the datasets which permits to track habits of particular user across mediums and their evolution over time.

4 Methods

Our work is divided into three main parts. At first, we proceed to realize a solely numerical analysis on the reviews using for this purpose the reduced datasets due to the performance limitations induced by the large ones. The second part consists of a natural language processing approaches applied to a pre-filtered set of reviews from the complete datasets. Eventually, we proceed to realize statistical test to verify the validity of the obtained results.

[1] `http://jmcauley.ucsd.edu/data/amazon/links.html`

4.1 Quantitative and Numerical Analysis

4.1.1 Data Processing

Even though, the datasets are consistent and well-formated thanks to the prior work of their authors, we need to remove some incoherent datapoints. Indeed, the Kindle premiered in October 2007 but some reviews are dated as being posted before this baseline. Hence we decide to only consider reviews starting from year 2008.

In addition, we notice that some users posted a suspiciously important amount of reviews in a given time span. We categorize such users as artificial reviewers[2]. We filter those by setting a threshold of 30 for the monthly maximum number of books and ebooks reviews.

In addition, we focus on users who consume books on both mediums and which have sufficiently many reviews on a monthly basis throughout the 2008-2012 period. We settle with a consistent yearly average of at least 0.5 books a month.

Eventually, we keep the reviews of the users who have reviews on both mediums books and Kindle ebooks and who conforms to the previously specified requirements.

4.1.2 Query Optimization

Given the size of the datasets used we decide to extensively use Pyspark in association with its MySQL implementation to enhance the speed of our queries.

4.2 Reviews Sentiment Analysis

After having extracted and analyzed the habits of the customers population who both frequently read books and ebooks, we investigate the difference in ebooks review sentiments of this users. We focus on the technical aspect of the Kindle support for this purpose. We use a Latent Dirichlet Allocation (LDA)[3] to determine if a "technological" topic can be observed among the review texts. Then, with the help of a variant of the TF-IDF word weighting, we search for the words more frequent in ebooks reviews and that fall into this category. We constitute a dictionary which allows us to separate two kinds of ebooks reviews *technical* and *non-technical* based on the presence of terms from the dictionary. Eventually, we apply the `TextBlob` sentiment analysis metric to the respective sets of reviews.

4.2.1 Reviews Pre-Processing

The first step to perform the desired analysis is to create a so called *bag of words* containing all the words constituting each review.

In order to improve the accuracy of the LDA and the TF-IDF variant, we apply a stemming step to the list of words constituting the reviews. This treatment yields the root of the words abstracting its specific grammar e.g *reading* turns into *read*. We use two different stemmers from the `nltk` library. In our case, `PorterStemmer` yields the best results for the LDA step and `Snowball` with 'English' as parameter provides us with the best results for the TF-IDF variant and sentiment analysis with `TextBlob`.

Eventually, to increase performance and precision further, we remove the top 1000 most usual words in the English language. Even though TF-IDF has a turnaround to this problem, this is not the case of the LDA algorithm.

4.2.2 LDA

LDA is a supervised learning method that permits extraction of topics from the corpus of reviews and to measure a similarity between reviews and topics. In the high level picture, LDA considers documents as mixtures of topics, themselves being considered as mixtures of words. The algorithm iteratively assigns words to topics.

Implementation is done with the help of pre-built LDA models present in Pyspark. The training of the model is realized on the stemmed corpus of reviews. Two hyperparameters α and β representing respectively the prior document and topic distributions are tuned according to our needs to yield the best results.

4.2.3 TF-IDF variant

In order to carry on the analysis started with the LDA, the technique we choose aims to select words that dont appear in a usual book review but that do appear in an ebook review. Hence a variant of the TF-IDF is applied between the corpus of the Kindle ebook reviews and the one of the classic book.

The variant we used is composed of the Term-Frequency matrix which schematically maps a word to the fraction of the words this word represent.

$$TF_{d,w} = \frac{N(d,w)}{|d|}$$

where $N(d,w)$ is the number of times the word w appears in the document d and $|d|$ is the size (number of words) of the document d.

The variant of the Inverse Document Frequency we used represents the proportion of time a word is supposed to appear with respect to the corpus, in our case it simplifies to :

$$IDF_w = \frac{1}{N}\sum_{i=1}^{N} TF_{D_i,w}$$

where N is the number of documents in the corpus. Note that this IDF is not the same as the usual used metric. This variation allows us to keep the words that are present in the entirety of the corpus.

Eventually, we can compute a variant of the TF-IDF matrix which is the combination of the two previously

[2] Independent, *Amazon bans biased reviews that have been influenced by brands*, 2016

[3] `https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation`

described matrices :

$$TFIDF_{d,w} = \frac{TF_{d,w}}{IDF_w}$$

This matrix has high values for words frequent in the given document compared to the rest of the corpus. Again some modification are added to the original formula to properly describe our problem.

4.2.4 TextBlob

We use the `TextBlob` method from the TextBlob Python library in order to obtain a convenient and simple sentiment analysis of the reviews text. The algorithm yields for each text a score in the continuous $[-1, 1]$ range where -1 corresponds to a very negative sentiment, 1 to a very positive sentiment and 0 a neutral sentiment.

4.3 Statistical Tests

In order to validate the results obtained in the text sentiment analysis step, we perform two statistical tests to verify these. Our null hypothesis is that the two sentiment analysis have the same distribution.

4.3.1 p-value Year Tests

We apply a p-value test to reviews aggregated yearly.

Let's call $T \sim \mathcal{N}(\mu_T, \sigma_T)$ the distribution of the sentiment for one technical review and $O \sim \mathcal{N}(\mu_O, \sigma_O)$ the distribution of the sentiment for one non-technical review.

Then we define $A = \frac{1}{N}\sum_{i=1}^{N} T_i \sim \mathcal{N}(\mu_A, \sigma_A)$, the average of the sentiment over the technical reviews.

And we can derive :

$$P(A \leq \mu_T | H_0) = 1 - \mathcal{Q}\left(\frac{\sqrt{N}(\mu_T - \mu_0)}{\sigma_0}\right) \quad (1)$$

The details of the full derivation of this formula can be found in part 2.4.1 of the notebook.

4.3.2 t-test on Similar Sample Sizes

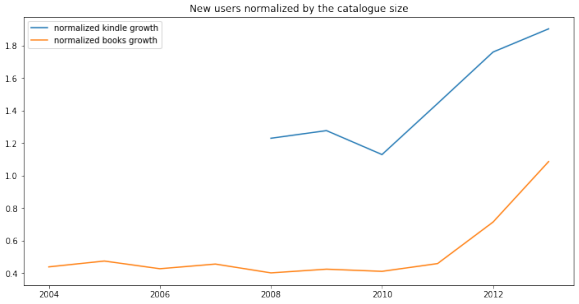
In order to abstract the difference in the number of samples for the technical and non-technical set of reviews, we decide to realize a t-test on samples of the same size for the two respective datasets following the removal of the outliers. We choose a sample of 10000 points for this purpose. This important number allows us to properly approximate the t-values with quantiles of the normal gaussian distribution.

The details along the mathematical model used can be found in part 2.4.2 of the notebook.

5 Results

5.1 Numerical Analysis

5.1.1 New Users Appeal

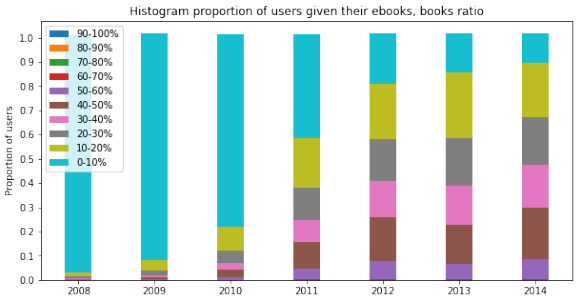


We observe that the ebooks appeal more to new users with respect of the size of the catalog of available titles. We remark for both books and ebooks an important increase in the number of new users which can be interpreted as a consequence of the general growth of the Amazon.com website in the last years.

5.1.2 Reading Pace

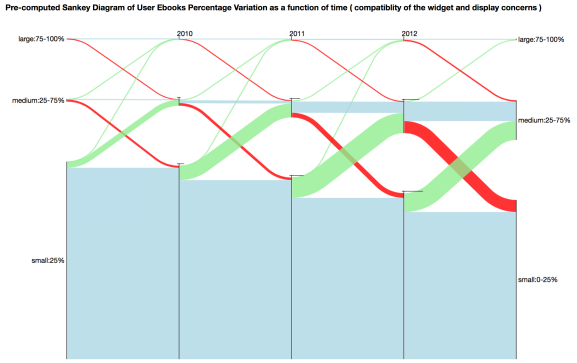
With the help of this plot, we can observe the impact of the kindle on the behaviour of frequent readers. We denote that the number of books regardless of the support remains relatively steady since the appearance of the kindle. However, we denote that the ebooks replace a proportion of about 1 out of 4 books read by the frequent readers. They compensate for the decrease of physical books reading. We notice that the ebooks tends to have a more important part in the reading ratio over years.

5.1.3 Ebooks-books Ratio Distribution



This more precise distribution displays the fact that the number of readers who read a larger proportion of ebooks tends to grow over time. However, the most represented ratios are still those under 50% of ebooks.

5.1.4 Kindle Ratios Variations



This sankey plot reveals the "migration" of users from a given ebook percentage to a lower or higher one. It splits the readers into three populations with respect to their ebooks to total books. We remark that in the last year, the proportion of users reducing their ebooks consumption is significantly bigger than the one occurring in 2008 and 2009.

5.2 Review Sentiment Analysis Results

5.2.1 Technology Lexicon Extraction

The analysis yields a Kindle-specific topic, containing (stemmed) words such as "kindl", "easi", "worth", "purchas" and "price", among all the literaly genres ones. However constructed topics are not significant enough for a meaningful reviews extraction.

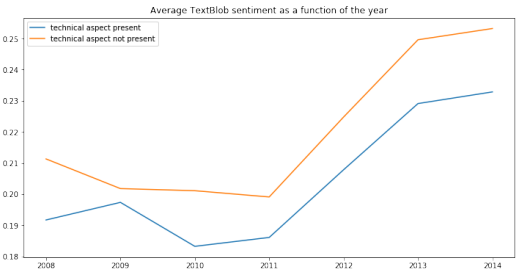
From the LDA analysis, we know that readers are indeed talking about some technology related topics. In order to extract those reviews from the whole set, we construct the required dictionary for filtering by analyzing the TF-IDF results at multiple levels (filtering words that appear once every 1 million reviews, then once every 100k reviews, ...) and removing every not platform-specific words.

charger	paperwhit	headphone	warranti	e-ink	well-format	recharge	tablet	kindl	ebook	e-read
1.965	1.94	1.883	1.877	1.777	1.775	1.6	1.587	1.574	1.551	1.535

download	screen	device	easi	complaint	price	disappoint	worth	recommend	hate
1.46	1.437	1.324	1.312	1.307	1.289	1.28	1.26	1.241	1.187

The specificity of the words displayed with respect to the Kindle corpus is displayed as a gradient from *green* for the most specific ones (values close to 2) to *red* for the less specific ones (values close to 1, meaning that this word is almost as frequent in the Kindle reviews than in their paper counterpart ones).

5.2.2 Sentiment Analysis Overt the Technology Aspect



We observe that the average TextBlob sentiment score is a little smaller for reviews containing the technical aspect that the ones which lack its in their review text.

In the next subsection, we present the results obtained for the statistical testing regarding this matter i.e. that the technical aspect has a negative effect on the sentiment score.

5.3 Statistical Test Validation

5.3.1 p-value Yearly Validation

Applying 1 to the yearly aggregated scores for each set of reviews technical and non-technical, we obtain the following p-values across the years.

	2008	2009	2010	2011	2012	2013	2014
p-value	0.0512	0.3012	3e-06	2.63e-13	2.01e-63	2.22e-193	4.73e-179
0.1 confidence	✓	✗	✓	✓	✓	✓	✓
0.01 confidence	✗	✗	✓	✓	✓	✓	✓
0.001 confidence	✗	✗	✓	✓	✓	✓	✓

The p-value for the hypothesis H_0 is way less than 0.001 for all the years after 2009. We even reach p-values that can be considered as null after 2012.

We still denote an outlying result in 2009 where the p-value is 0.3012 . However, computing the the p-value the other way around (making the same computation under hypothesis H_1) also gives us a high p-value of 0.3165 . Hence this result is as probable under hypothesis H_1 as under hypothesis H_0. Therefore the result obtained in 2009 should not be a reason to keep the null hypothesis.

5.3.2 Time Invariant t-test

Applying the statistical test setting described in section 4.3.2, we obtain a value of $t = 9.00046$ which indicates a high degree of confidence in this second analysis.

Hence, given the two statistical test carried, we affirm that we can safely **reject the null hypothesis** and consequently that the distribution of the reviews sentiment regarding the technical aspect of the Kindle is not distributed in the same fashion as the one for the reviews about the non-technical aspects.

6 Conclusion

We observe that the Kindle appeals to a lot of new users who try this new format and its popularity is important given its several orders smaller catalog compared to classical physical books. However, frequent readers do not switch over time to the new platform. The consumption of traditional physical books is still the most significant amount those users. In addition a decreasing tendency in the ebooks ratio is observed in the last period. This can be interpreted with the recent trend among younger generations to favour old kinds of media supports.(Sian, 2017)

The sentiment analysis part of our work shows that users tends to have a small negative bias towards the kindle reading support. We confirm with statistical tests that this difference although being small is statistically meaningful and not due to randomness among the samples.

References

[Merga2015] Margaret Merga. 2015. *Do Adolescents Prefer Electronic Books to Paper Books?*, *Publications, 3(4), 237e247.*.

[Lee, Jolles, Lamonato2017] Pierre-Alexandre Lee, Marc Jolles, Yves Lamonato. 2017. *Books versus eBooks : The customers choice*.

[McAuley2017] R. He, J. McAuley. 2016. *Modeling the visual evolution of fashion trends with one-class*, *WWW, 2016*.

[Sian2017] Sian Cain 14.03.2017. *Ebook sales continue to fall as younger generations drive appetite for print, the Guardian. https://www.theguardian.com/books/2017/mar/14/ebook-sales-continue-to-fall-nielsen-survey-uk-book-sales*.

A Figures and Plots Appendix

For convenience, the reader can find all graphs and figures in high resolution on the github repo https://github.com/xavierpantet/ada-project-1 in the folder `figures`.