How to choose phone's brand according to Amazon

Marion Kramer
marion.kramer@epfl.ch

Alexandre Rassinoux
alexandre.rassinoux@epfl.ch

Kristina Satara
kristina.satara@epfl.ch

Abstract

Reviews help when making decision on which product to buy, as they give possible customers personal experiences and more product details. Since customers usually invest some time to study the reviews and ratings before buying specific products from Amazon, the idea of this project is to create evaluation platform which would help the users to choose their future product. For that, insight into downsides and advantages of various brands based on the reviews will be given. The aim of this project is to analyze reviews of different brands of Amazon cell phone products, and to decide with the help of the reviews to decide which features are important for specific brands of cell phones. For this natural language processing techniques, such as Naive Bayes Classifier and Latent Dirichlet Allocation will be used.

1 Introduction

Sentiment analysis, also known as opinion mining or emotion AI, is important part of natural language processing. The subject of the study is the attitude or the emotion behind certain texts. The purpose of sentiment analysis is to determine weather text content is neutral, positive or negative towards some topic.

Buying a cell phone on Amazon is not an easy task. The website does not provide useful information on cell phones features like battery, screen or apps. Most of the time, customers have to read a set of reviews in order to have an idea on the product, and sometimes reviewers posted counter opinions.

The aim of this project is to provide to customers an overview of phone brands characteristics based on the reviews.

Data used for this project is collected from Amazon, during the period of few years. Most of the data is collected between 2012 and 2014, as shown in Figure 2. Each of the reviews has also ratings that are used for ground truth. The ratings are based on 5-star system, where 5 is highest and 1 is lowest rating. Distribution of the ratings is J-distribution with highest number of 5-star reviews, and it is shown in Figure 1. After purchasing some of the products, customers are asked to give their review and rating, which are further subject to sentiment analysis in our project.
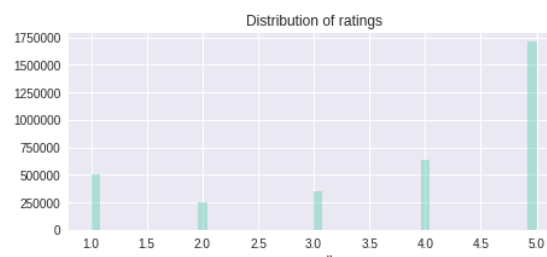

Figure 1: Distribution of the ratings in reviews

Figure 2: Number of reviews during time

2 Related work

Natural language processing is an interesting and popular area of research, thus there are many papers regarding this topic. As Amazon is a widely used platform, it has gained much reviews in the past few years. Focus was made on the cell phone reviews, and papers which gave different approaches and conclusions were found. Most of them are using techniques like bag of words and

n-grams for sentiment analysis. There is also significant number of papers that are using machine learning techniques for this. Here can be found few examples of papers that are interesting and important regarding this project. In paper *"Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning"* by Callen Rain, the approach is using Naive Bayes Classifier and bigrams for tagging the words for deciding about important features. Another relevant paper in this area of research is *"Deep Learning for Amazon Food Review Sentiment Analysis"* by Jiayu Wu and Tianshu Ji. They gave different approach to the similar problem by using Recurrent Neural Networks for sentiment analysis.

3 Data Collection

The datasets used for this project are provided by Julian McAuley. It consists of two json files - one containing the reviews, and another providing metadata about the products. These two files were merged for further data analysis. The data has also been transformed in order to improve the speed of loading and processing the data in a python notebook. As data was incomplete, a decision was made to enrich it with additional dataset from Kaggle platform. This is public GSMArena dataset and it contains more than 8000 cell phone specifications.

4 Dataset Description with Summary Statistics

After merging and combining all the datasets, some descriptive statistics have been made. Given below are the information provided in the datasets:

- Reviewer's ID
- ID of the product reviewed
- Name of the reviewer
- Text of the review
- Rating given
- Summary of the review
- Time of the review
- Price in Euros
- URL of the product image
- Brand name
- Category of the product

Firstly, an exploration of the brands of cell phones producers that the dataset contains has been made. As expected, most of the brands are known ones. The interest was given into calculating the mean and median prices of these cell phones. Figure 3 gives overview of mean prices for 25 most popular brands.
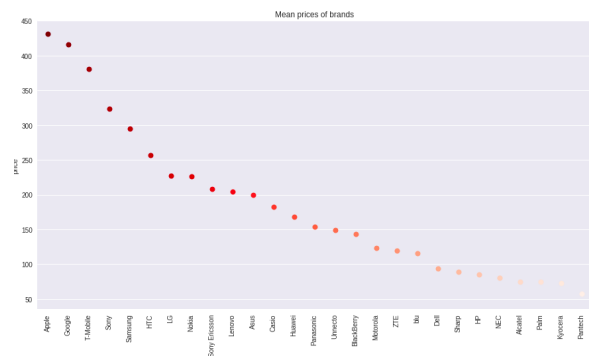

Figure 3: Mean prices per brand

Have a look at price distribution can be interesting in order to understand better the dataset and specific models contained in it. It is shown in Figure 4.

Figure 4: Distribution of the prices

Another statistics which is interesting is average number of cell phone models per brand, mean ratings per brand, and models that have highest number of reviews.

Customer's decision about purchasing specific model is often influenced by overall rating of the product - but in the project it has been shown that this rating should be taken very carefully when used for making decisions. This is not the rating of specific model that is important, but also the number of reviews. During the exploratory analysis, it was found that some models only have a few reviews, but their rating is the highest possible, and there are other models that have an average rating above four and also have few thousand times more reviews. This should be taken into consideration very carefully before the decision.

There are no correlations between the price and the rating. Figure 5 shows example of correlation

of words with the ratings - either high or low rating for specific brand, in this case it is *Sony Ericsson*. It can be seen that some words have higher correlation with the reviews - for example battery and phone.
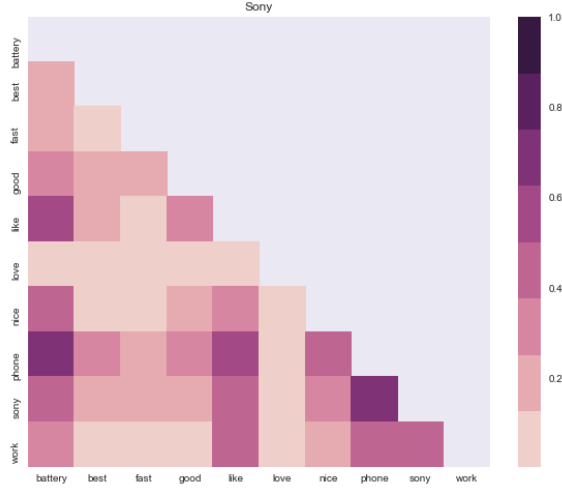

Figure 5: Correlations of the words with reviews

This was the basic analysis of the dataset. Further analysis is closely connected to the natural language processing and sentiment analysis, thus it will be explained into more details in the next section.

5 Methods

Two approaches were used in order to find the important features of cell phone brands. The first approach is to use supervised learning to classify new review as positive or negative. The second one is the use of topic modelling and sentiment analysis techniques in order to get an opinion based on most important phone features.

5.1 Naive Bayes Classifier

The reviews are split into positive and negative based on the ratings, and they are separately analysed. Decision was made that ratings four and five will be positive reviews, and one and two will be negative reviews. Reviews rated with three were not taken into account because they were considered as neutral reviews. After splitting reviews into positive, negative and neutral groups, supervised learning was used to check if it was possible to train the model in order to predict the sentiment of some new reviews. Data was firstly pre-processed - tokenized, and stopwords are removed. After that, lemmatization is done in order

to merge the different forms of the same word into one, by removing affixes.

The model used in the project is Naive Bayes Classifier, which applying Bayes theorem with the *naive* assumption of independence between every pair of features. In this particular case, features are word tokens. Given a rating ω_j and an dependent word token x, Bayes theorem states the following relationship:

$$P(\omega_j|x) = \frac{p(x|\omega_j) \cdot P(\omega_j)}{p(x)} \quad (1)$$

90% of the data were used for training and 10% for the validation, and with Naive Bayes Classifier, a result of 84% of successful classification was obtained. Most important features for classifying the review as negative are words *scam* and *false*, whereas for classifying the review as positive important words are *perfect* and *excellent*. As the aim of this project is not to find only the important words in reviews, but to determine which of some specific cell phones features - like battery, ergonomy, screen, camera, etc - are correlated and important for some brands, a second step is needed to go further in the analysis.

5.2 Further Sentiment Analysis

Here a sligthly modified approach was used. First, the reviews were split according to the brand and according to the ratings - five groups are created for ratings from one to five.

A topic modeling step is done using words vectors and Latent Dirichlet Allocation (LDA). This algorithm is used to discover the topics that are present in the set of reviews, and it is based on probabilistic graphical modeling. The set of reviews can be seen as a bag of words, and topics as a distribution over a fixed vocabulary. Words are generated from document specific topic distributions. This step enables to get 10 words per topics, and 10 topics per rating, for each brand.

The second part of the work is to attach a sentiment tag on topic words. The Vader (Valence Aware Dictionary and sEntiment Reasoner) analysis technique is implemented in order to get the weighted sentiment score of a word. This method produces four sentiment metrics from these word ratings. The first three, positive, neutral and negative, represent the proportion of the text that falls into those categories. The final metric, the compound score, is the sum of all of the lexicon ratings.

Here, two kind of words have been considered. First are feature words that are where the most interest is - like battery or screen. Second group of words are non-feature words (great, bad, problem, work, etc). Only non-feature words in topics will be subject to sentiment analysis.

In a specific topic, the sum of weighted scores of all non-feature words will be given to feature words. If a topic does not contain feature words, it is ignored. At the end, the results are grouped to get an overview for all the reviews and all ratings. The density of appearance for each feature word in each topic is counted. As a result, the sum of polarity scores for each feature word is computed.

6 Results and Findings

From the exploratory analysis the next conclusions can be made: The years when there were the higher number of reviews are 2013. and 2014. Overall the most reviewed models by the customers are *Samsung Galaxy S3 Mini, Nokia N8, Motorola Moto G*. Best rated cell phones are not always the best choice, as they sometimes have only few reviews. There are no correlations of the prices and ratings. Based on our dataset, the higher mean price of cell phones has *Apple* with average price of almost 450 euros.

Regarding the sentiment analysis, for each brand there are usually more positive than negative reviews. As explained in the previous part, using the bag of words and Latent Dirichlet Allocation and analysing the important feature words, the results obtained are shown on Figures 6 and 7.
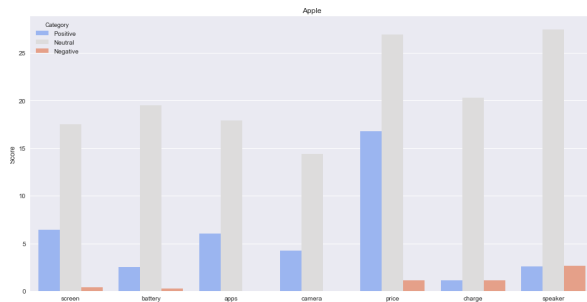

Figure 6: Sentiment metrics for cell phone features

7 Conclusion

This platform uses reviews and extracts meaningful information from it, and then evaluates certain cell phone brand based on these reviews. Senti-
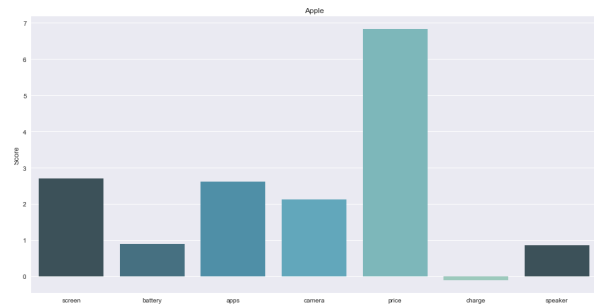

Figure 7: Compound score for cell phone features

ment analysis was used to classify each of the reviews and it's parts as good or bad. After this process, appearances and sentiments have been found for the feature words. For the review preprocessing, the next techniques are used : tokenization, removal of the stopwords, lemmatization.

For classifying the reviews, Naive Bayes Classifier is used, while for detecting the feature words, bag of words and Latent Dirichlet Allocation were used. Based on this, specific brands and features that customers find important for certain cell phone brands are described and evaluated.

References

[wallin] Alexander Wallin. *Sentiment analysis of Amazon reviews and perception of product features*.

[Liu-Bing] Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies.

[Callen Rain] Callen Rain. *Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning*. Swarthmore College.

[Hutto] Hutto, C.J., Gilbert, E.E. June 2014. *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. Eighth International Conference on Weblogs and Social Media (ICWSM-14), Ann Arbor.

[Julian McAuley] J. McAuley, C. Targett, J. Shi, A. van den Hengel 2016. *Image-based recommendations on styles and substitutes*. WWW.

[Elli-Wang] Maria Soledad Elli, Yi-Fan Wang. *Amazon Reviews, business analytics with sentiment analysis*. Department of Computer Science, North Carolina A&T State University, USA.

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

[Fang Zhan] Xing Fang and Justin Zhan. 2015. *Sentiment analysis using product review data* . Department of Computer Science, North Carolina A&T State University, USA.