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, we decided to create evaluation platform which would help the users to choose their future product. For that, we are giving insight into downsides and advantages of various brands based on the reviews. 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 reviewms to decide which features are important for specific brands of cell phones. For this we use natural language processing techniques, such as tf-idf matrix and Latent Dirichlet Allocation.

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.

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. We first use Naive Bayes Classifier, training it with one part of the data, and testing it on the another part. We gained 84% of correct classifications when classifying the reviews as positive or negative. After this we use bag of words and Latent Dirichlet Allocation (LDA) for extracting interesting features for the brands.

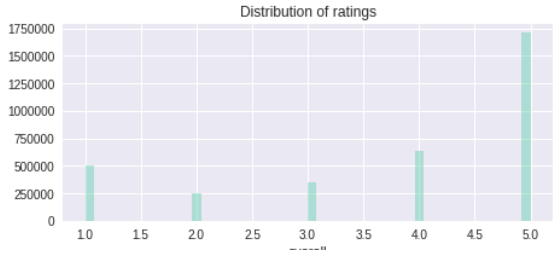

Figure 1: Distribution of the ratings in reviews

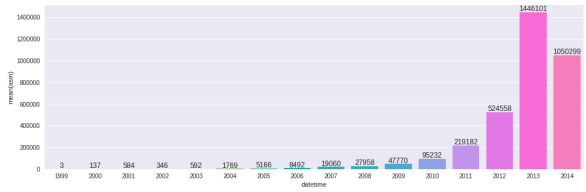

Figure 2: Number of reviews during time

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. We focused on the cell phone reviews, and we found papers which gave different approaches and conclusions. 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 tech-

niques for this. We will give here few examples of papers which we find interesting and important. 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 diffent approach to the similar problem by using Recurrent Neural Networks for sentiment analysis.

3 Data Collection

We used dataset provided by Amazon for this project. Dataset consists of two json files - one containing the reviews, and another providing metadata about the products. We merged these two files for our further data analysis. We also transformed the data in order to improve the speed of loading and processing this dataset in our python notebook. As this dataset was incomplete, we decided to enrich it with additional dataset from Kaggle platform. This is public GSMArena dataset and it contains more than 8000 phones specifications.

4 Dataset Description with Summary Statistics

After merging and combining all the datasets we had, we did some decriptive statistics. 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

We already had some expectations, so we first explored brands of cell phone producers that our dataset contains. As expected, most of the brands were known to us. We were interested in mean and median prices of these cell phones. Figure 3. gives overview of mean prices for 25 most popular brands.

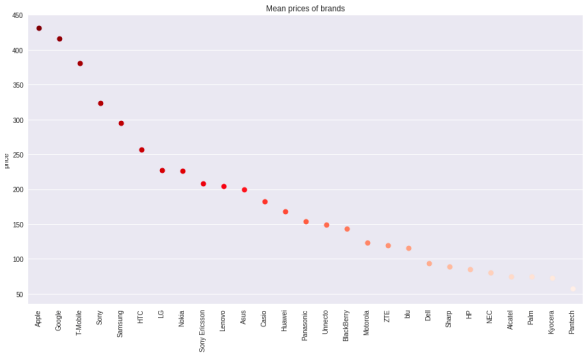

Figure 3: Mean prices per brand

In order to understand better our dataset and specific models contained in it, we were interested in price distribution. It is shown in Figure 4.

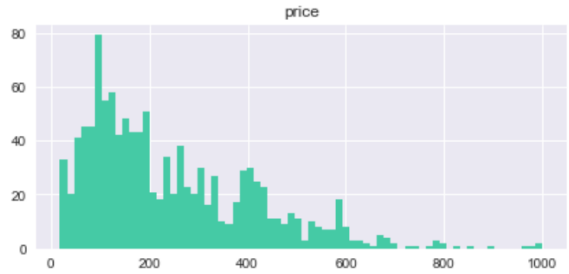

Figure 4: Distribution of the prices

Another statistics which we found interesting was average number of cell phone models per brand, mean ratings per brand, models that have highest number of the reviews.

Customer's decision about purchasing specific model is often influenced by overall rating of the product - but in the project we show that we should be very careful when using this metric for making the decision. Not just the rating of specific model is important, but also the number of the reviews. We found during exploratory analysis that there are some models that have only a few reviews, but the rating is highest possible and there are models with average rating above four which have few thousand times more reviews. This should be taken into consideration very carefully before the decision.

We found that 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. We notice that some words have higher correlation with the reviews - for example battery and phone.

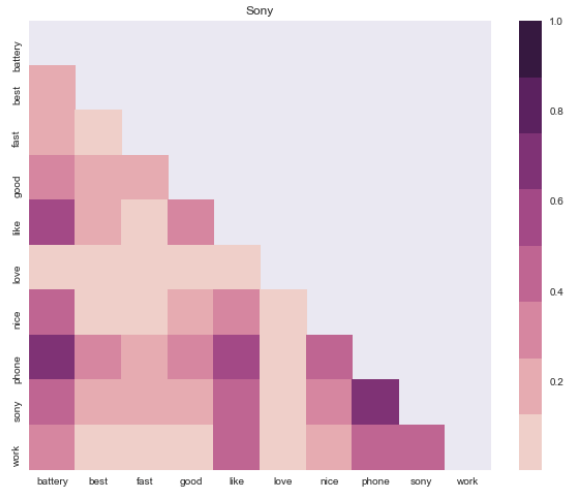

Figure 5: Correlations of the words with reviews

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

We used two approaches for finding important features of cell phone brands. The first approach is to use supervised learning to classify new review as positive or negative. After this we extract important words and do the sentiment analysis for the review using bag of words approach and Latent Dirichlet Allocation.

5.1 Naive Bayes Classifier

The reviews are split into positive and negative based on the ratings, and they are separately analysed. We decided to consider ratings four and five positive reviews, one and two as negative reviews. Reviews rated with three were not taken into account as we considered them to be neutral. After marking the reviews as positive, negative or neutral, we used supervised learning to check if we will be able to train the model to predict if some new review is positive or negative. The reviews are first tokenized and stopwords are removed. After this lemmatization is done in order to merge the different forms of the same word into one, by removing affixes.

We used 90% of the data for training and 10% for the validation, and with Naive Bayes Classifier we got result of 84% of successful classification. 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 our project is not to find only the important features, but to determine which of some specific features - like battery, ergonomy, screen, camera, etc - are correlated and important for some brands, we go further with the second step in the analysis.

5.2 Further sentiment analysis

Here we take sligthly modified approach. We undertake the next steps:

- First we split the reviews according to the brand and according to the ratings - five groups are created for ratings from one to five.
- Next we do topic modelling using words vectors and Latent Dirichlet Allocation (LDA). Reviews are seen as a bag of words, and topics as a distribution over a fixed vocabulary. Words are generated from document specific topic distributions. As the output of this step we have 10 topics with 10 most important words for each topic (for each brand).
- After this, we tag a sentiment on topic words. We use the previous Sentiment Analyzer to get a weighted sentiment for the topic. Here we consider two kind of words. First are feature words which we are mostly interested in - like battery or screen. Second group of words are non-feature words (great, bad, problem, work, etc). Non-feature words in the topic will be subject to sentiment analysis.
- The weighted scores of all the non-feature words will be given to all the feature words in the following topic. If a topic does not contain feature words, it is ignored.
- At the end, we group the results to get an overview for all the reviews and all ratings. We count the density of appearence for each

feature word in each topic. As a result and we show the sum of polarity scores for each feature word.

6 Results and Findings

From the exploratory analysis we can make next conclusions: 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 choise, 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 we explained in the previous part, using the bag of words and Latent Dirichlet Allocation and analysing the important feature words we obtained the results shown on the Figure 6. TODO: ALEX INSERT THE NICE FIGURE HERE AND ADD FEW SENTENCES HERE TO CONCLUDE PLEASE :)

7 Conclusion

This platform uses reviews and extracts meaningful information from it, and then evaluates certain cell phone brand based on these reviews. We use sentiment analysis to classify each of the reviews and it's parts as good or bad. After this process, we find appeareances and sentiments for the feature words (for example battery, screen, camera, etc). For the review preprocessing are used next techniques: tokenization, removal of the stopwords, lemmatization.

For classifying the reviews Naive Bayes Classifier is used, while for detecting the feature words we use bag of words and Latent Dirichlet Allocation. Based on this, we describe and evaluate specific brands and features which customers find important for certain cell phone brands.

References

Callen Rain *Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning* , Swarthmore College

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

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

Liu, Bing *Sentiment Analysis and Opinion Mining*, Synthesis Lectures on Human Language Technologies

Alexander Wallin *Sentiment analysis of Amazon reviews and perception of product features*