Hearing the customer's voice – sentiment classification and summarization of Amazon reviews

Stefan Memmer (May 2020)

Springboard Data Science Career Track

Executive Summary

What clients feel or think is key to improve products and to increase sales

Relying on datasets from interactions with customers can be helpful to „hear“ what customers really think, appreciate or dislike

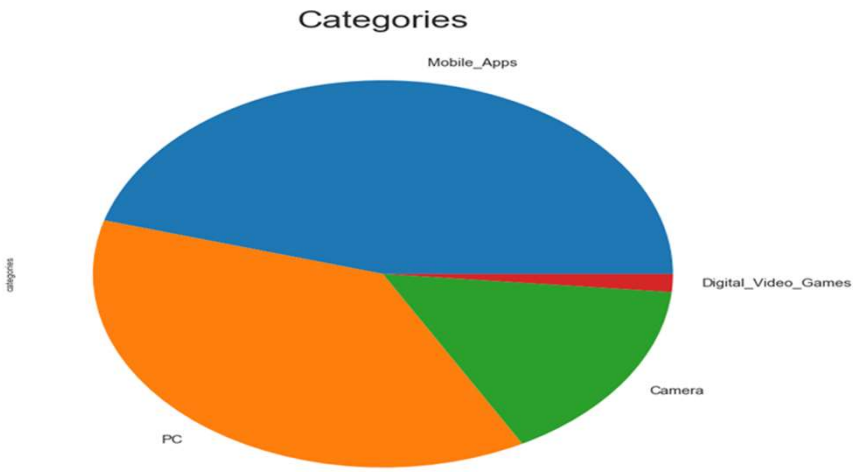
Unstructured datasets like emails or Q/As are hard to handle and cumbersome to label

With the example of Amazon product reviews an approach gets presented which is able a) to generate sentiment classifications and b) to extract summaries by leveraging Deep Learning

First steps on Data

The dataset is generated by randomly mixing reviews about Mobile Apps, PCs, Cameras and Digital Video Games:

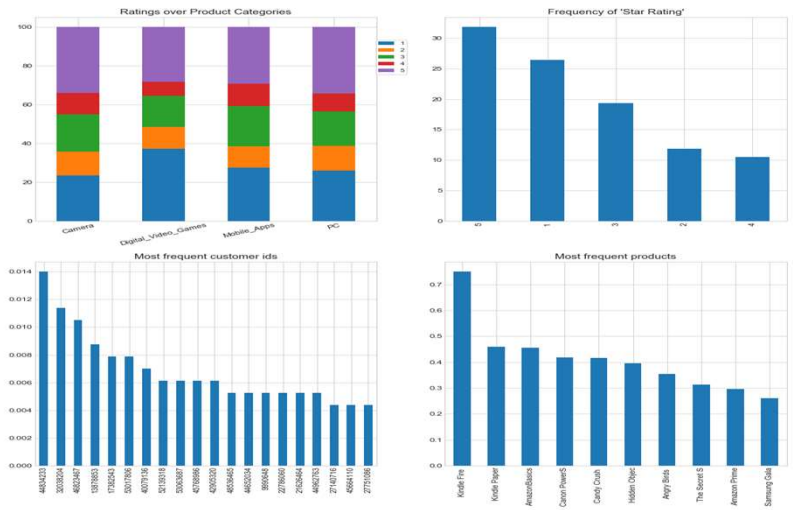
The final size of set is ~114,000 reviews with following product proportions:



Further steps on data & Exploratory Analysis

Cleaning included dropping duplicated reviews and those with NA values

Besides text, the reviews have more descriptive features:



- First chart right: distribution of star ratings over product categories
- Second right chart: overall frequency of star ratings
- First lower chart: percentage of unique customers
- Second lower chart: Most frequent products

Data Exploration – features

Strong bias towards polarity - 5 star ratings most frequent

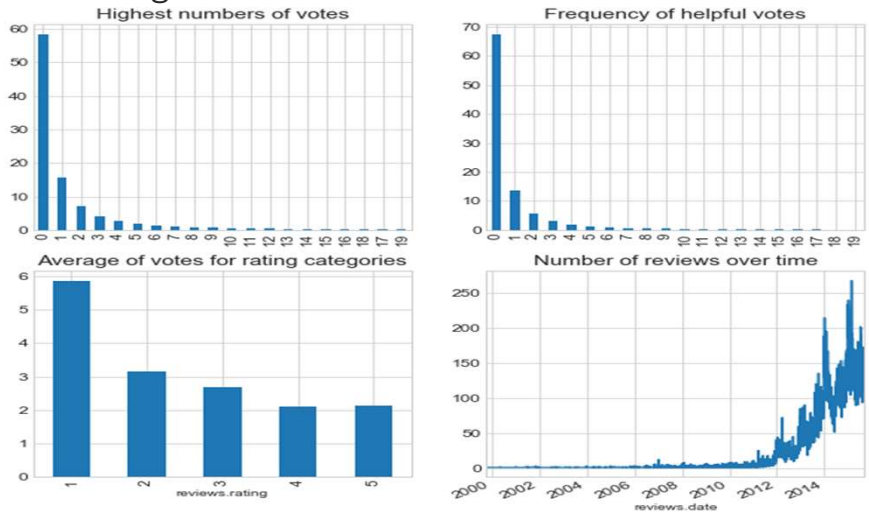
Most reviews are about Kindle Fire, followed by Kindle Paper..

Digital Video Games show comparably high proportion of negative ratings

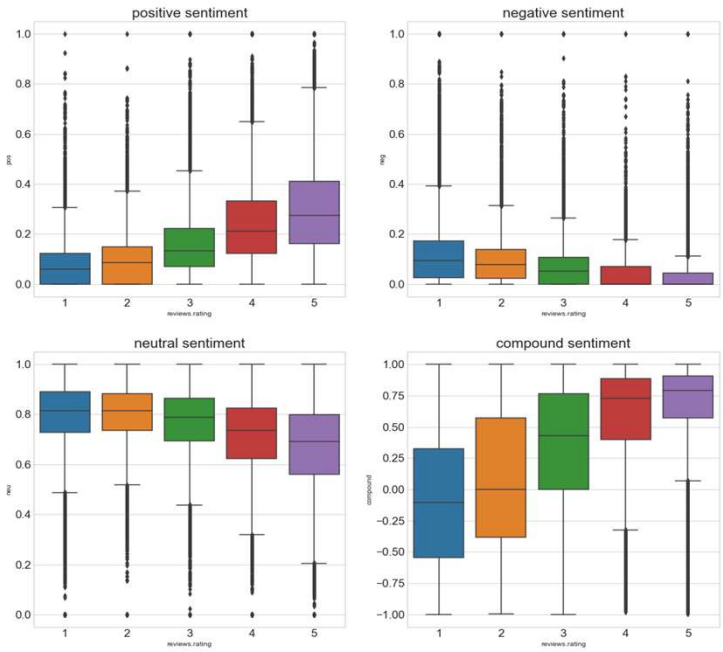
Most reviews have not been voted by others…

only those with bad ratings (bottom left chart)

since 2011, number of yearly reviews increases

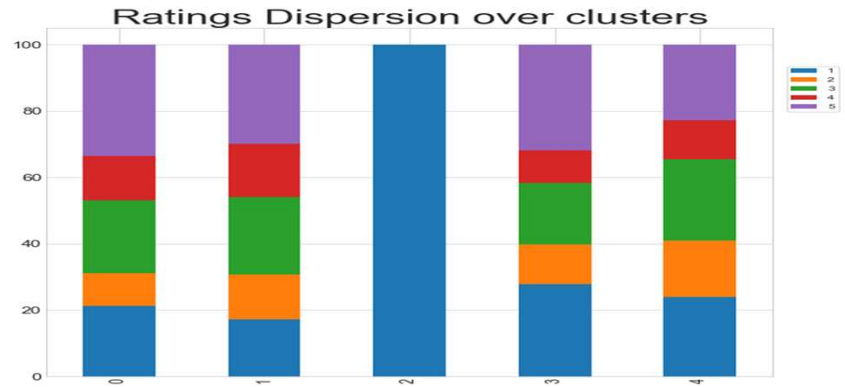


Data Exploration – reviews‘ text



- Overall understandability of reviews corresponds to pre-colllege level („ARI score“ of 17.00)
- Vader‘s lexical approch of sentiment scoring yields distributions across reviews with different star rating (left figure)
- Higher „neutral sentiment“ for badly rated reviews hints at possible behavioural bias, i.e. neutral verbal expressions despite dissatisfaction

Data Exploration – review's text



- Word Cloud shows most frequent words of the dataset
- Clustering reviews:
  - Basic tokenization for document-term frequency matrix
  - NMF to extract clusters and topics
- Rating frequency over different clusters:
  - Cluster 0 & 1 show overall positive ratings
  - Cluster 2 – entirely negative ratings
  - Cluster 3 & 4 – higher 1 & 2 Star frequency

Deep Learning Approch for Classification & summary extraction

First, further cleaning by dropping too short reviews and those containing many irregular expressions

Language Model: pre-trained Deep Learning model to generate word encodings

Classfication Model: based on learned word embeddings a DL classifier network gets trained to distinguish sentiment poles, i.e. (4&5 star – 1 &2 star) and neutral (3 star ratings)

Extraction of representative reviews via clustering embeddings

Language Model

Pre-trained ULMFit model is trained further on subset of the reviews to gain context specific word embeddings

LSTM structure of encoders to predict word given the previous one

Learning rate finder and cyclical learning rate scheduling

epoch	train_loss	valid_loss	accuracy	time
0	3.782862	3.689017	0.304888	00:52
1	3.622302	3.583864	0.312590	00:52
2	3.498814	3.540415	0.316205	00:52
3	3.382692	3.530978	0.317090	00:52
4	3.203698	3.547635	0.315948	00:51
5	3.043352	3.583271	0.313255	00:52
6	2.908148	3.624678	0.311251	00:52
7	2.754258	3.670273	0.307939	00:52
8	2.663227	3.702912	0.306286	00:52
9	2.609444	3.718244	0.305493	00:52

- After 10 epochs 30 % of time the correct words were predicted
- one word out of 33,000 words
- Major target to generate word embeddings that represent contextual and semantic information of words used in reviews

Classification Model

The language model‘s last layers are changed to linear layers suitable for categorical classification

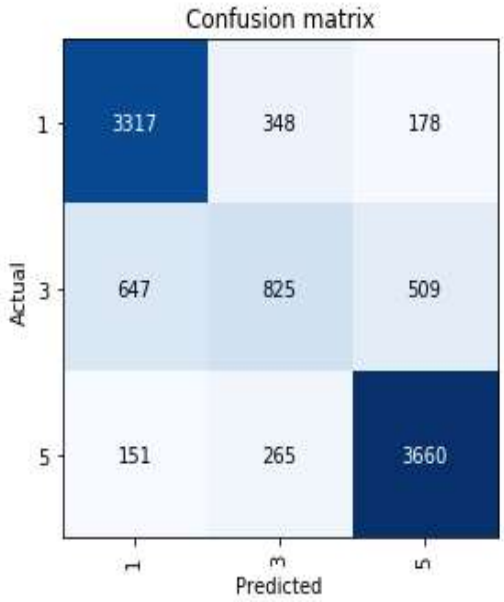
The language model‘s word embeddings are transferred to classification model

Training set 33,000 reviews with 1/4 as validation set

Test set 12,000 reviews

Process of gradually unfreezing layers in the network and reducing respective learning rates

Classification results on test set



	precision	recall	f1-score	support
1	0.81	0.88	0.84	4191
3	0.60	0.42	0.49	2069
5	0.85	0.90	0.87	4324
accuracy			0.80	10584
macro avg	0.75	0.73	0.74	10584
weighted avg	0.78	0.80	0.79	10584

- overall accuracy of 80 %
- Most failures with the middle (neutral) category
- Reason: behavioural bias of reviewers, i.e. writing style

Extracting summaries

Approach:

- Averaging word embeddings to generate vector representations of reviews
- Clustering the reviews‘ aggregated embeddings by K-Means
- Number of clusters either given by user or square root of review matrix as default
- Reviews which are closest to respective cluster center are selected as most representative one

Provides possibility to summarize review text on pre-defined subsets

Cluster technique ensures selected reviews to be different, at least at the level of learned embeddings, and thus, to provide distinct infromational content

Conclusion

Approach shows certain advantages

- Classification and Summary Extraction in one stroke
- Transfer Learning in NLP skips off the need of large datasets and domain specific word/sentence representations are achievable
- Deep Learning might be better used to capture „fine grained“ nuances of human language

For product/client managers this approach helps to extract relevant and meaningful information about what customers think or desire

… without the need of huge datasets or spending large efforts to label datasets containing interactions/feedback from clients