M3 EXAM – STAKEHOLDER REPORT

Michael Dybdahl

Aalborg Universitet
September 2019

Description of data acquisition

I found the dataset [Women's E-Commerce Clothing Reviews] from kaggle which contain 23k reviews on women clothes bought online. This is real commercial data which has been anonymized, and references to the company in the review text and body have been replaced with "retailer".

Problem statement

How can the sentiment of a review be predicted using NLP, deep learning and non-neural baseline models?

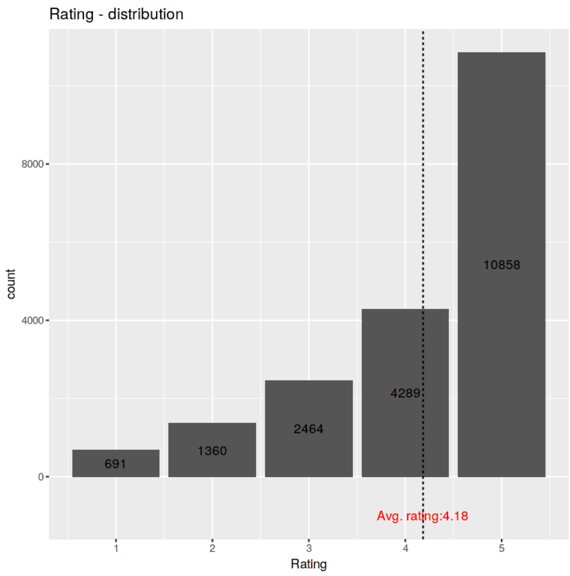
Having all these reviews from customers who have bought the clothes online, it would be interesting to see if I can build some models that can predict whether a review has a negative or positive sentiment, especially the negative reviews can be interesting to look at - to see what the firm can improve.

I will use NLP to make a tidy format of the data and find the sentiment of the reviews.

1. Exploratory data analysis (EDA)

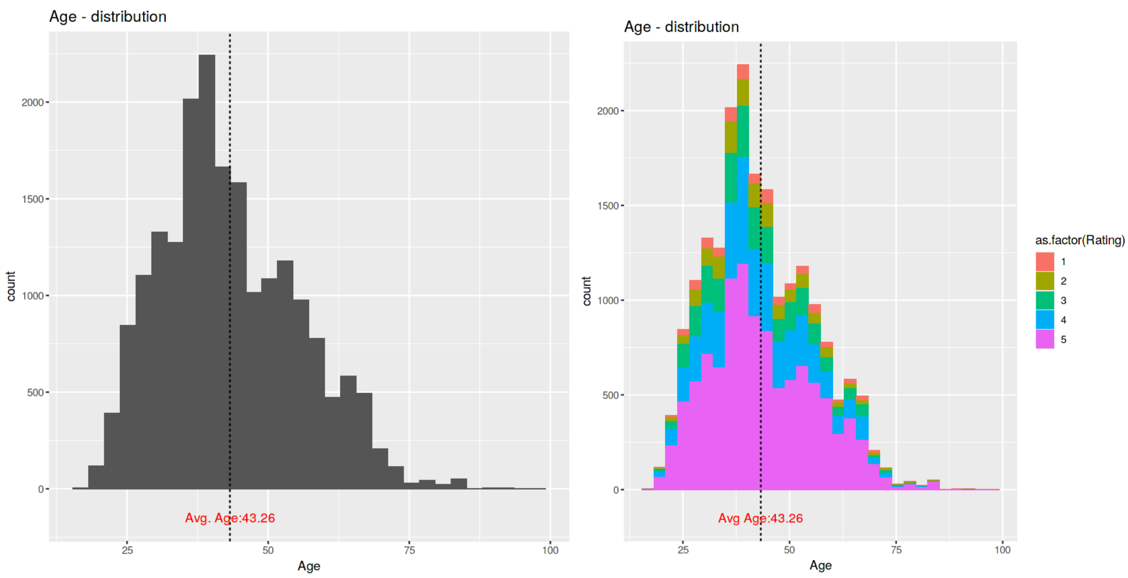
First, I did some EDA to look depper into the data before I began modelling.

Rating:



As you can see, then the highest rating (5) accounts for over half of the distribution, and the 2nd most counted rating is rating 4. We can also see that the average rating is 4,18, which is pretty high for a scale from 1 to 5.

Age - distribution:



We can see that the average age is 43,2 year and the distribution is pretty normal distributed. This make some sence, because probable the young girls do not think about reviewing their purchases online, and perhaps not many older women shop online.

Not much to say about the age distribution and ratings, because it seems mostly the same among the age distribution.

3. Data preparation

First, we **cleaned** the data for variables that we were not going to use and removed reviews with NA values.
After this we recoded some of the variable's names and the format of "data" and "rating".

2. Natural Language Processing (NLP)

2.1. Bag-of-words (BoW)

First, I created a bag of words, where this tidy format where I going to use for both adding the sentiment for each review and the non-neural baseline model.

The top words where "dress" and many positive words, which makes sense because of the high rated reviews.

2.2. Sentiment analysis

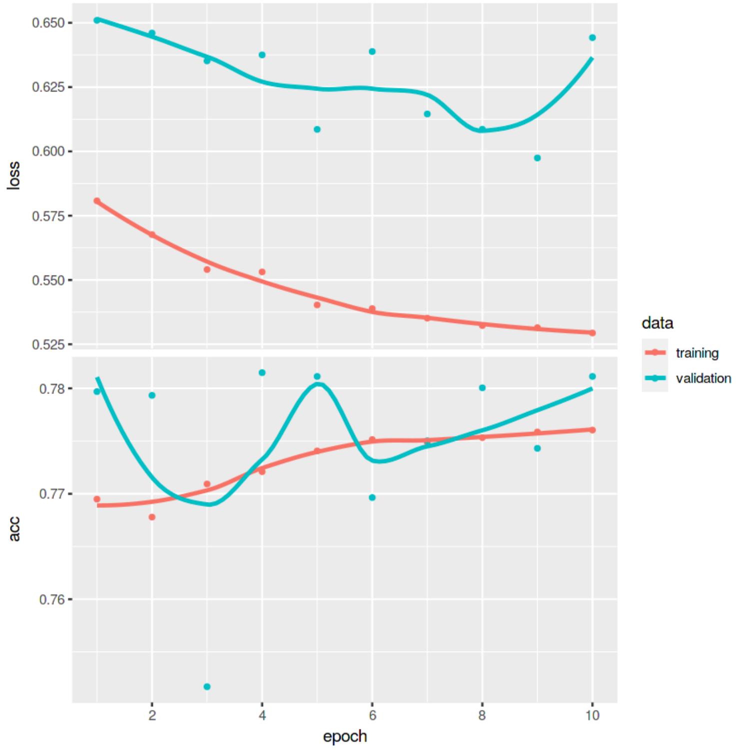
With the tidy dataset, I used the "bing" lexicon to add the sentiment (negative/positive) for each review.

3. Deep learning models

Before running some models, I splitted the data into a training and test set and preprocessed the data, so it was ready for training and running some DP models.

3.1. Basic feed-forward ANN

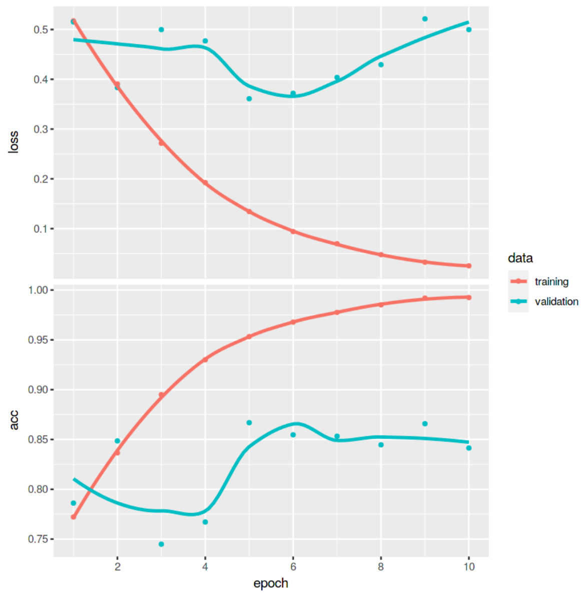
I started by making a basic feed-forward ANN model.



Evaluation:
The model performed not that good with a accuracy at 77,5%, but it only predicted the true positive reviews and almost none true negative reviews (the underrepresented class).

3.2. Simple RNN model

I also ran an RNN model with a simple rnn and embedding layer.

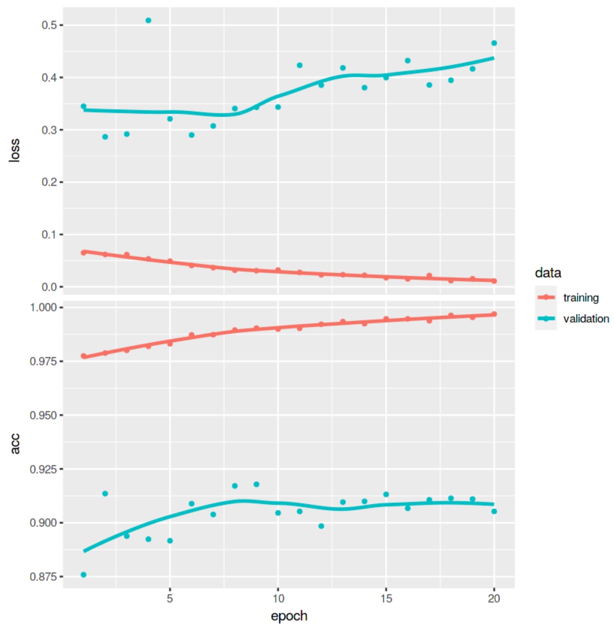


Evaluation:

The model performed significantly better than the basic FF model, having an accuracy at 84,9%. It also predicted the negative reviews 67,03% correct, so much better.

3.3. LSTM model

At last I ran a LSTM model.



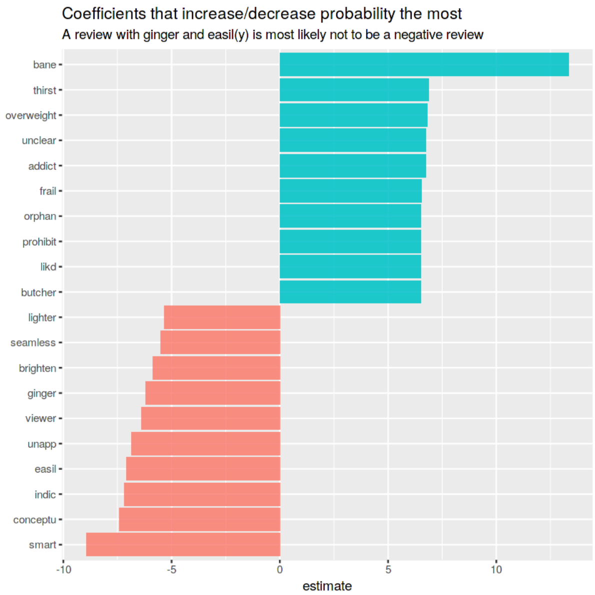
Evaluation:

The LSTM model performed much better than the previous models, having an accuracy at 90,93%, although "only" predicting 68,9% negative reviews correct.

4. Non-neural baseline model

After the DP models, I did a non-neural baseline model using tidy data.

Again, I was trying to predict whether a review is Negative or Positive according to the sentiment.



Evaluation:
This model performed pretty good, with an accuracy at 97,84% - predicting 94,21% negative reviews correct.

The cause of the much better performance could be that I used the tidy dataset, which had more processing (removal of stop words etc..) than the data in the deep learning models.

5. Conclusion

Looking at the dataset "Women's E-Commerce Clothing Reviews" with around 20k reviews, I did some NLP and added the sentiment to the reviews, which I tried to predict using DP and a non-neural baseline model.

The **deep learning models** performed fine, where the LSTM model was the best one with a accuracy around 91% and predicting the underrepresented class, Negative reviews 69% correct.

The **non-neural baseline model** performed even better with a accuracy at 97,84%, where 94,2% of the negative reviews was predicted correct. The higher accuracy can probably be caused by more and better preprocessing if the tidy data, where the DP models only used preprocessing tools from keras.

Perhaps the **deep learning models** performance could have been improved by experimenting more with the architecture and tuning, but sadly I ran out of time + GPU Quota for kaggle (30 hrs) was used and first reset after the hand-in.

Overall what could I use the results for? Well predicting the sentiment of reviews, look into negativ reviews and what words affect the sentiment, which can be used for companies to see what have a negative impact on their product reviews.

Further development could be to look deeper into what products and categories thathave the most negative reviews, divisions. Maybe also the reviews with high feedback counts, because this would probably be the most useful reviews.