Amazon Alexa Reviews Sentiment Analysis
Machine Learning for Natural Language Processing 2020

Auger Jean-Baptiste
ENSAE Institut Polytechnique
Jean.Baptiste.AUGER@ensae.fr

Abstract

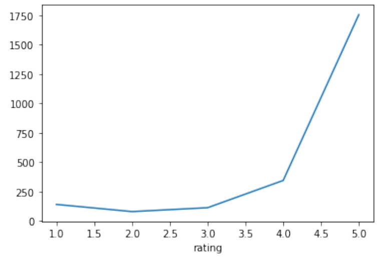
Using Amazon's reviews and ratings, we compare two models for sequence labelling. The first uses binary labelling, whereas the second uses a multi-class classification which better captures the distribution of the ratings. We found that both models are accurate when predicting a 5 rating. However, the second model hardly differentiates between the other ratings [1].[2]

1 Problem Framing

The use of ratings of a product has grown exponentially in the last few years. They represent large amounts of data where customers leave a written review and a rating to the product. Even though this could entail problems such as coherence between ratings (as people with approximately the same comment might give a different rating based on personal preference), the existence of such data is an opportunity to develop a sentiment analysis where the sentiment is directly correlated to the rating. We use data from the Amazon Alexa review to develop a sentiment analysis. We aim at extracting quantitative metrics from the reviews based on a sentiment analysis, and we try to predict the sentiment of other reviews. More precisely, we ask ourselves if we should analyse sentiment with the scale provided by the rating system, or by a binary option : "positive" and "negative", in the case of a asymmetric use of the rating system. To test this, we compare a binary model and a multi-class classification model.

2 Experiments Protocol

We use data extracted from Amazon's website that contain 3150 Amazon Alexa reviews. The data can be found here. The ratings are skewed, as 5 is overly represented, whereas 0,1,2,3 and 4 are less represented. This might entail some prediction's errors in the end.



We divide the data in three subsets : train, test and dev.

The vocabulary is quite extended, as there 7733 different words used. After tokenization the nltk package, we could reduce the vocabulary size up to 5213.

We built two sentiment analysis models based on BERT. They both use the same preprocessing that tokenizes the reviews and adapt them to the BERT framework. The loss function is the mean of the sum of the labels predicted times their associated log probability. We then evaluated our models both qualitatively and quantitatively using different metrics such as f1 score, accuracy, precision, recall. The first model we trained is a binary sentence classification. It takes advantage of the particular distribution of the ratings, which is largely skewed towards five. The second model we put in place is a multi-class classification model. It takes advantage of the fact that there are 5 different ratings possible.

3 Results

The first binary model, on the one-hand, has a f1_score of 0.9, a loss of 0.39, a precision of 0.89. As a result, we consider that this model is able to

[1] https://colab.research.google.com/drive/1-dV9KJIqAOlQ$_aEDbmgQNG_Go54tiN6kscrollTo = g4O - vXJmy6BC$

[2]

differentiate reviews whose authors rated Alexa 5 from the others. We also tested this model qualitatively by asking him to predict the label of several sentences. He assigned a value ”1” (which is positive) to sentences such as ”Alexa is very helpful”, whereas sentences such as ”it’s ok I guess” would receive a label 0.

The multi-class model, on the other hand, had a macro average precision of 0.52 and a weighted average precision of 0.75. The difference can be explained by the fact that there has much more 5 ratings than the others. The f1_score is 0.42 on average and 0.75 when considering weighted average. The precision of the model is below 0.5 for the labels ”3” and ”4”, which can be explained by the sample size. We also tested qualitatively the model on the same sentences that we used on the first model. The model was able to spot the difference between ”It’s ok I guess” and ”Alexa is very helpful”, giving respectively the labels 4 and 5. However, it gave a label 4 to tis sentence ”That was absolutely awaful”, which is an oddity.

4 Discussion

Our analysis has shown that a multi-class neural network was not able to capture the sentiment that triggers a rating other than five. This could come from two main issues. The first is the fact that a very small portions of the rating are different from five, which means we have little data to analyse the sentiment between the rating 1 to 4. Plus, the very meaning of those rating may vary grandly between individual. As a result, the distinction between a 1 and a 2 may change radically depending on the person rating the product. As for future work, we would suggest to enhance the multi-class approach, maybe by getting a wider database that could get different reviews. This would need to be validated from a statistical point of view, as population might be different, which could induce difference in the tone of the reviews and the rating.