Sentiment analysis on cultural products in Amazon Review data
Machine Learning for Natural Language Processing 2022

Lasme Ephrem Essoh
DSSA

Mathieu Chabaud
DSSS

Nathane Berrebi
DSSA

Abstract

In this project we are interesting in cultural products and their evaluation by Amazon customers. We realize a sentimental analysis on digital music products and try to transfer it to other types of cultural goods to see whether these products are comparable. Our sentimental analysis on the digital music products works quite well, not only with a LSTM model but also with a BERT model. However, our attempt to transfer the prediction on movie and TV products is not successful due to different types of vocabulary. Finally, we also explore a sequence generating task by trying to summarize the reviews in a few words.

1 Problem Framing

For the last 20 years there has been a renewal of the economic thought concerning the particularities of cultural goods. According to (Karpik, 2007), cultural goods like music, movies or books are very specific economic goods since their quality is by essence subjective. We might think therefore that each type of cultural goods has its own codes to judge the quality of a product, which cannot be reduced to assessing whether a product "works" or "doesn't work". This hypothesis of an imperfect transposition of the judgments of the quality of products among different types of cultural goods can be partially verified by exploring the commentaries that people make on the cultural products they buy online. In this prospect we use the Amazon Review dataset to test whether predictions we make on the sentiment a customer had on a given type of cultural product (here digital music) are useful to analyse sentiment on another type of cultural product (books).

2 Experiments Protocol

In this project we make a supervised sequence labeling task. Indeed, the Amazon review dataset provides for each product a text review corresponding to the comment made by the customers along with the rate he/she gave to the product. This settings allows to have a direct correspondence between text data (the review) and a label corresponding to the sentiment of the customer on the product (the rating). To illustrate this correspondence, we show hereafter the most common bi-grams of words in the review according to the rating (positive, neutral and negative) for the digital music dataset:

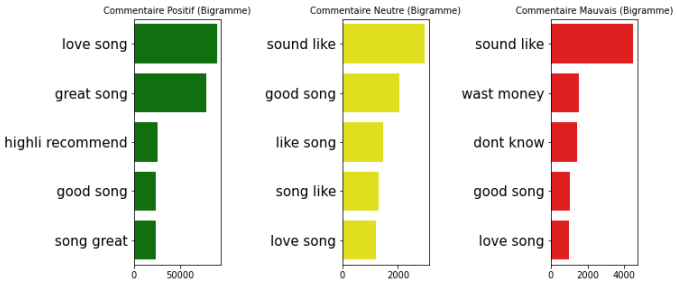

Figure 1: Bigrams of words according to the rating

We can see that in positive reviews, the more frequent bigrams are "love song" and "great song" ; fore neutral reviews, the more frequent bigrams are consistently more neutral ("sound like" and "good song") ; finally, for the negative reviews, the more frequent bigrams are "sound like" and "wast money", this one being logically a very negative appreciation.

Our experimental protocol consists first in training a basic LSTM model to predict 2 classes (positive and negative reviews) on digital music products. Before that we preprocess the data by removing stop words and punctuation and by using the Porter algorithm ((Porter, 1980)) to take the stem of the word. Then we take a subsample of our data to avoid too long computations and to balance the data. Indeed, we have originally a very un-

balanced data with much more positive than negative reviews. Then, we train a LSTM model with dropout layer to avoid overfitting. Note that we do not use pre-trained embeddings (like fastText for example) to let the neural network learn itself the best embeddings according to the vocabulary we have in the data. We take as usual the accuracy as the evaluation criterion of the model. We save the best model (on the test set) and we will consider it as our benchmark for the rest of the project. Then we take data on another type of products (movies and TV) and we evaluate the best model we trained on digital music to see how it performs on other type of data. Second, we try another modeling approach with BERT, which differs from our benchmark LSTM model in 2 ways: first, it has already been pretrained on very large data ; second, it has specifically been designed to solve sentiment analysis problem. Because this algorithm is known to have better performance, we consider here 3 categories of reviews: negative, neutral and positive. We still evaluate the results on the accuracy criterion. Finally, we try at the end of our project to explore an other task of NLP which is summarization. For this, we predict the summary of a review from the review. This task of sequence generation is still a supervised task because we have in our dataset both reviews and summaries. Since this part was only exploratory because it demands a lot of time to train, we do not evaluate it quantitatively but only qualitatively.

3 Results

Our baseline binary classification LSTM algorithm reaches an accuracy of 80% on the digital music data. This is quite satisfactory given that we trained it only on 20,000 observations to avoid too long calculations. However, when we tried to apply the model on other type of data (movies and TV products), it failed to give satisfactory results and reached only a 56% accuracy. This might be due to the fact that the vocabulary of movies and TV reviews was different, so the weights learned in the digital music data were not directly transferable to other data.

Our BERT estimation also provides satisfactory results, with an accuracy of 75% (based also on a 20,000 observations sample). This is worst than our baseline model, but it makes sense since we allowed for an additional class (neutral) of ratings. Moreover, we tried to run the algorithm on a larger sample, and it resulted on a better accuracy (we reached 92% of accuracy after 23 hours of training).

```
luckily introducing katie thiroux doesnt make the listener choose this is truly the best of both worlds.
Suggestions de titres: {'it is the album to me', 'ive bought this album', '<|endoftext|>'}

anything that this man has ever sang was a hit i wish i owned every song he ever sang even his movie would be great to own really though i
love listening to gospel as much as a love old country.
Suggestions de titres: {'Love it.', 'Very Good Music', 'This man has always made me want to own every song that he ever sang.'}

okay.
Suggestions de titres: {'Best Five Stars', 'Very good', 'Five Stars'}

this item was just what i need at the time it arrived in a timely fashion and was well packaged.
Suggestions de titres: {'Very pleased', 'I like this item', 'Very Awesome item'}
```

Figure 2: Summarization of reviews

Finally, our first attempt to do a summarization task provided interesting results. Even with only 1,000 observations, we were able to generate summaries which make sense, as illustrated below. However, this task would be better implemented with more data and with a rigorous evaluation metric.

4 Discussion/Conclusion

The interest of this project was first to evaluate the performance of sentimental analysis on cultural goods, which might not be as evident as in other classes of products which have inherent properties which make them good products or not. The subjective nature of the reviews on cultural goods is not necessarily an obstacle to efficient sentimental analysis since people can be attached to give very good comments (resp. very bad comments) when they like (resp. dislike) a song or an album. We suggest indeed that sentiment analysis task can be performed with a satisfactory precision on these types of goods. However, we were unable to clearly assess whether it is possible to transfer the analysis of reviews from one type of cultural goods to another. This is due to the fact that the vocabulary of the reviews are different according to the product. A direct improvement would consist in retrain our model on pre-trained vocabulary to have similar embeddings for all the products. We were lacking of time to do it but it will be according to us the best thing to do. Another improvement would consist in train our summarization algorithm on much more data to have a better idea of its performance.

References

M F Porter. 1980. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137.

Lucien Karpik. 2007. *L'économie des singularités*. Gallimard.