Challenging BERT and its variants on Sentiment Analysis Task
Machine Learning for Natural Language Processing 2021

MBIA NDI Marie Thérèse
Student
marietherese.mbiandi@ensae.fr

VONIN Cédric
Student
cedric.vonin@ensae.fr

Abstract

This work was realised for the Machine Learning for Natural Language Processing class. The aim of this study is to Compare the results of BERT model with its variants on the sentiment analysis task. For the sake of comparing models, we use use three variants of BERT: Roberta, general Character BERT and medical Character BERT. Our approach relies on proposing an efficient treatment for unbalanced class in order to facilitate the learning during the training of models. At the end, we have stated that BERT outperforms its variants on sentiment analysis task.

1 Problem Framing

BERT, which is a new language representation model, stands for Bidirectional Encoder Representations from Transformers, it was introduced by Google in 2018 (Devlin et al., 2019). Afterwards many variants of BERT were developed, some on specialised vocabularies and others re-trained versions of BERT. These variants performed better than BERT on specific tasks (a Masked Language Modelling task (MLM) and a Next Sentence Prediction task (NSP)) and specific datasets. The natural question would be to ask whether the improvement remains on other tasks with new datasets. This is the main objective of this work. We will compare the performance of BERT with three of its variants on sentiment analysis task, using transfer learning. For this sake, we will use RoBERTa, general characterBERT and medical characterBERT. RoBERTa (Yinhan Liu, 2019) is a fine tuned version of BERT, it performs better than BERT on MLM and NSP, so we can assume that RoBERTa might perform better than BERT on sentiment analysis. CharacterBERT (Hicham El Boukkouri, 2020) is similar in every way to BERT but uses a different method to construct initial context-independent representations: while the original model consults its vocabulary to split unknown tokens into multiple wordpieces then embeds each unit independently using a wordpiece embedding matrix, CharacterBERT uses a Character-CNN module which consults the characters of a token to produce a single representation. This model is robust on misspellings and noise, this is why we chose it. Medical characterBERT and general characterBERT are the characterBERT trained respectively on medical and general corpora.

2 Experiments Protocol

2.1 Dataset

To perform the sentiment analysis we use the dataset of Amazon Fine Food Reviews provided by Kaggle. The dataset is available on this link. The dataset has 568454 rows and 10 columns. We are only interested in comments and ranks so, we will only keep two columns of the dataset (score and Text). The rank is between 1 and 5, so we classify the rank into negative (if the rank is between 0 and 2), neutral (if the rank is equal to 3) and positive (if the rank is higher than 3) sentiment.

We choose to no remove HTML tags, stopwords, punctuation because we want to keep the structure of each comment,since we will be more comfortable with detecting implicit negative comments. It's common that users duplicate theirs comments and score to avoid time loss. So, let's check if it's the case with this dataset. If we have duplicate comments, we keep the first comment among duplicated comments. The new dataset has 395200 rows and we split it on train (75%), test (12.5%) and evaluation (12.5%) sets.

The weakness of this dataset is to be unbalanced. Indeed, as shown in figure 1, the classes were not uniformly distributed among the dataset.

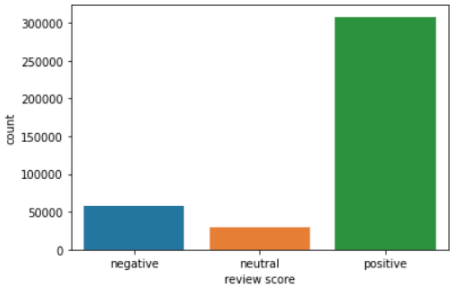

Figure 1: distribution of sentiments

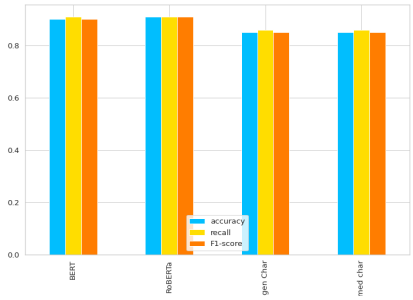

Figure 2: qualitatives results

To tackle this issue, we use data augmentation techniques provided by this paper on train dataset. Due to computatioal resources, we were able to augment only 2000 reviews on class 0 and class 1.

2.2 Training

CharacterBERT being a very big model, our computational power permitted us to train it on only 3% of the data. For this task, we used the BERT base cased model because of our desire to consider accent and case (it is usefull to understand the review). The hyperparameters used during the training phase were those indicated as optimal by the authors of the models. To deal with the issue of unbalanced classes, we weight every class in the loss,as follows: 0.3 for class 0, 0.5 for class 1 and 0.2 for class 2. We used the same optimizer (AdamW) for all models and the same loss function (Cross Entropy Loss). BERT and RoBERTa were trained on 4 epochs and both characterBERT models on 15 epochs as they have less data.

3 Results

In this section, we present our experiments' results. For quantitative comparisons, we will use accuracy, recall and f1 score. As we can see on the figure 2 that RoBERTa has better results than BERT (the difference is not too high). characterBERT models have have the smallest results and same results.

It is interesting to see how these models perform on the least represented class. RoBERTa has the best result as F1-score on the class 1 is 53% followed by the BERT with 47% on F1-score. Character BERT model had bad results on this class (about 30%). This bad performance can be explained by the amount of data used during training (3% of the train data). With these quantitative results, we can state that RoBERTa and BERT perform better. So we will use only those two models for qualitative evaluation. For qualitative evaluation, we use financial dataset. The target here was to evaluate the most flexible model on foreign vocabulary (different from the training vocabulary). This dataset is provided by kaggle and avalaible here. So we use a neutral, positive and negative reviews. RoBERTa predicts all the reviews as positive reviews and BERT predicts neutral and positive reviews as positive and negative review as negative. You can look at our notebook for more details. Both models were not able to detect the neutral review and predicted it as a positive one. In this task, RoBERTA has worse results than BERT as it was able to give a true prediction on one comment. With our experiments, we can say that BERT has better results than the Three other models on sentiment analysis.

4 Discussion/Conclusion

To conclude, this project was the opportunity for us to compare BERT results with its variants on sentiment analysis. although we have noticed that, in this task, BERT outperforms its variants, the performance of characterBERT models suggest that they would have performed better than BERT if they had had enough data. One of the major problem is having to deal with the computational costs. The performance of our computers was limited so, we were not able to perform more data augmentation and to train characterBERT model on the whole training set. Moreover, we were not able to fine tune the training (we used hyper parameters provided by authors). Therefore, a future work will be to allow more computational ressources to fine tune models, do more data augmentation and use 2 or 3 datasets.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805v2*.

Naman Goyal Jingfei Du Mandar Joshi Danqi Chen Omer Levy Mike Lewis Luke Zettlemoyer Veselin Stoyanov Yinhan Liu, Myle Ott. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692v1*.

Thomas Lavergne Hiroshi Noji3 Pierre Zweigenbaum Junichi Tsujii Hicham El Boukkouri, Olivier Ferret. 2020. Characterbert: Reconciling elmo and bert forword-level open-vocabulary representations from characters. *arXiv:10392v3*.