# *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 has been 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, which stands for Bidirectional Encoder Representations from Transformers, was introduced by Google in 2018. 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 is actual 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 the sentiment analysis task, using transfer learning. For the sake, we will use RoBERTa, general characterBERT and medical characterBERT. RoBERTa 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 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 choose it. Medical characterBERT and general characterBERT are the characterBERT trained respectively on medical and general corpora. So we make another asumption: Does general characterBERT perform better than medical CharacterBERT on foreign vocabulary (not medi- cal).

## 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 classsify 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,so we will 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 and just keep the first comment among duplicates comments. The new dataset 395200 rows and we split it on train (75%), test (12.5%) and evaluation (12.5%) sets.

The weakness of dataset is to be unbalanced. In-

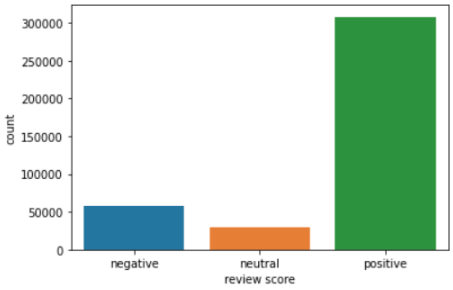


Figure 1: distribution of sentiments

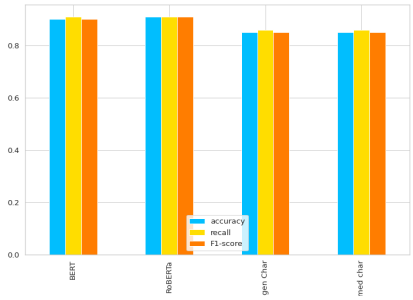


Figure 2: qualitatives results

deed, as shown in figure 1, the classes are not uniformly distributed among the dataset. 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 is a very big model with our computational ressoucres we was not able to train it on the entire dataset so, we use 3% of data. For this task, we use the BERT base cased model because we matter about accent and case (it is usefull to understand the review). The hyperparameters used during the training phase were those indicated as otpimals by the authors of the models. To deal with the issue of unbalanced classes we weight every class in the loss, so 0.3 for the class 0, 0.5 for the class 1 and 0.2 for the class 2. We use 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.

Next, we will compare result of BERT with RoBERTa and general characterBERT with medical characterBERT because of the shape of training set on characterBERT models.

## 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 RoBERTa have a better results than BERT (the difference is not too high), characterBERT models have same results.

The interesting way is too see how these models performs on the least represented class. RoBERTa has the best result as the 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%) due to the among of data used during training. With these quantitative result we can state that RoBERTa and BERT outperform. 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. This dataset is provided by kaggle and avalaible here. So we use a neutral, positive and negative reviews. RoBERTa predicts all the reviews as poositive 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 no able to detect the neutral review and predict it as a positive one. In this task RoBERTA has worse result 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 has been the opportunity for us to compare BERT results on sentiment analysis with its variants. At the end we have found that in this task BERT outperforms than its variants. 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 tuned the training (we use hyper parameters provided by authors). Therefore, a future work will be allow more computational ressources to fine tuned models, do more data augmentation and used 2 or 3 datasets.

## References