Do vegetarian meals taste bad ?
Machine Learning for Natural Language Processing 2021

Clément Montes
3^{rd}-year student at ENSAE Paris
clement.montes@ensae.fr

Pauline Roubeix
3^{rd}-year student at ENSAE Paris
pauline.roubeix@ensae.fr

Abstract

We classify the comments on recipes and analyze whether more positive or negative sentiments are associated with vegetarian recipes. We find that comments associated with vegetarian recipes carry more negative sentiments than the meals which are not.

1 Problem Framing

Vegetarian and vegan people represent 5% of French population. Since France is the country of gastronomy, this project aims to know if vegetarian recipes taste as good as their meat/fish counterparts.

We base our study on a dataset of recipes from the website food.com found in Kaggle[1]. We label vegetarian recipes thanks to the tags of the recipe. We end up with 1 132 367 observations, 16% of which are from vegetarian recipes.

We construct the sentiment of every commentary based on the rating associated with the comment. We exclude 0 because the comments associated were not relevant. We gather grades of 1 and 2 to create negative sentiment (encoded 0). The grade of 3 will be associated with neutral sentiment (encoded 1). Grades of 4 or more are associated with positive sentiment (encoded 2). The grouping of grades came from our overview of the dataset. Though, the approximation is fairly good for positive and negative sentiments, we aknowledge a grade of 3 shows neutrality at the aggregated level, but while looking comment at a comment, it is a rough assumption.

One challenge of our project is to deal with that unbalanced dataset and the huge weight of the dataset.

[1] Link to the Kaggle source

2 Experiments Protocol

2.1 Word2vec and logistic regression

The embedding of our comments is as follow: removing of caracters due to the scraping, tokenizing comments, stemming during lemmatizing. We intentionally kept stop words. We then fit a Word2Vec model. The parametrization is mainly inspired by the practicals but also the Google's recommendations (for the dimension). The final column of embedded words corresponds to the mean of vector of words in the comment. That way we only have one vector *per* comment.

We stratify our train_test_split to deal with the unbalanceness. Unfortunately, the dataset was too huge and we decided to delete 80% of the comments with positive sentiments. The unbalanceness may induce bad predictions, but using SMOTE was not a choice in a Google Colab while dealing with a 3GO dataset.

We decided to run a logistic regression based on this github which indicates it works better for such a project. We then fit a multinomial logistic regression, optimize hyperparameters. We set weights for each class to deal with unbalanceness.

2.2 Bert

BERT (Bidirectional Encoder Representations from Transformers) is a bidirectional transformer architecture. A transformer is an attention mechanism that learns contextual relations between words (or sub-words) in a text. Bert is trained as a Masked-Language model. In other words, if you mask a word in a sentence, Bert should be able to predict it seeing the context (left and right context because it is bidirectional). In this part we want to use Bert to do classification for a sentiment analysis.
To do the classification, we needed a pre-trained model for Bert Sequence classification (we found

which one here). Before training this model on our dataset, we had to tokenize the reviews with a tokenizer already implemented in the transformers library. Then we chose to train the model during 5 epochs, but the dataset was too unbalanced so the training did not work at all and the model only predicted the comments to be positive (just as a naive classifier). At the end of the day we deleted 95% of positive comments and rerun the model.

3 Selecting the best model

While looking at the colour code of confusion matrices (Figures 1 and 5), it appears that BERT seems to be a better classifier. However, since we have several labels, it is interesting to see in fraction of true labels (Tables and 1 and 3). That way, it appears that the logistic regression is biased toward the positive comments (functioning kind of like the naive classifier). Here we want to see if negative comments target more vegetarian meal, therefore, we are particularly interested in the recall in label 0, in Tables 2 and 4 : once again BERT is way better. Eventually, while looking at the F1-score in every class, we see that BERT is better.

At the end of the day, BERT may be a better classifier because of the different datasets we used to train our algorithms. On the one hand, 80% of positive comments were dropped to train logistic regression meanwhile 95% were dropped in the train set of the BERT presented in Appendix. We did so because while running BERT in the same dataset than the one for logistic regression, it did the naive classifier. Moreover, in the notebook we can see that Bert overfitted the training set, but the results remain good on the test set.

As Bert seems to have better results, we will use this model for our analysis of the opinion of people on vegetarian recipes against non vegetarian recipes.

4 Results

Finally, with Bert we have a model that reaches a F1-score of 88% for the positive class (see table 3), which is rather satisfying. The fact that the F1-score is better for the positive class seems logical because it is the class with more observations.
Now if we take a look on the qualitative results, we can see that there are several reasons why supposed positive comments (with rating 4 or 5) are misclassified and said to be negative by our Bert model, and the opposite way. First, if we look at the examples in the notebook, we have that comments were associated with a good rating but brought a very negative message. This can be due to an error from the person who wrote the comment. Other comments are misclassified because they brought a nuanced message, with positive and negative aspects, so the algorithm is lost. Finally, there are rough mistakes from our model, with comments that say for example ”Excellent recipe” but are qualified as negative.

Our main question remains: are vegetarian recipes less appreciated than non vegetarian ones ? If we take the test set and the mean of the predictive sentiment values, we have that the mean of non vegetarian recipes is 0.22, and the mean of vegetarian ones is 0.20, so it seems that non vegetarian recipes are prefered. This is confirmed by the histogram of the three sentiments (see Figure 6).

5 Discussion/Conclusion

Our results defer from our point of view but we identify several extensions to the project. First, the logistic regression is estimated by maximum of likelihood. Since our dataset is unbalanced, the estimator is not consistent. One extension relays in the paper by King and Zeng (2001) which corrects the bias for rare events (which are negative comments in our case). The construction of our dataset corresponds to their, but we would have to estimate the average number of negative comment in cooking websites to plug it into the bias correction. Another important limitation is the fact that we don’t have a dataset with comments labelised as ”positive”, ”negative” or ”neutral”. We chose arbitrarily the labels thanks to the ratings but sometimes those labels are wrong, especially for the neutral class. There also could be a bias due to the facts that vegetarian recipes could be commented by all kind of people whereas non vegetarian ones may only be commented by non vegetarian people.

To conclude, we managed to construct classifiers that have possible improvements, but deliver interesting results given the bias we identified throughout that report. The main result is that vegetarian recipes seem to taste less good, but we think this dataset is full of haters :)

.

References

Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political Analysis*, 9:137–163.

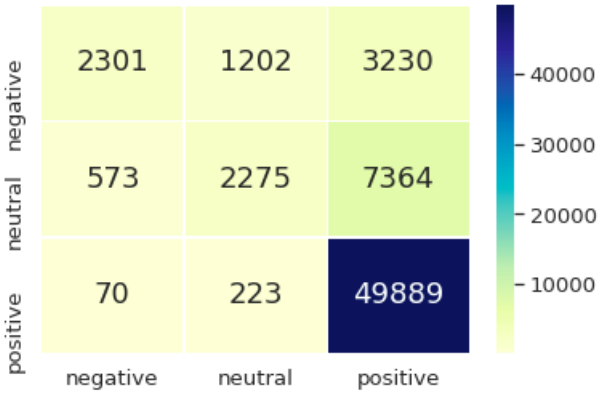

Figure 1: Confusion matrix of logistic regression
Note : x-axis is predicted labels. y-axis is true labels.

	Negative	Neutral	Positive
Negative	.34	.18	.48
Neutral	.06	.22	.72
Positive	.001	.004	.99

Table 1: Confusion matrix of logistic regression in percentage of true labels
Note : x-axis is predicted labels. y-axis is true labels. Computed manually from the above matrix. We find recall on the diagonal.

Label	Precision	Recall	F1-score
0	.78	.34	.48
1	.61	.22	.33
2	.82	.99	.90

Table 2: Classification report of the logistic regression

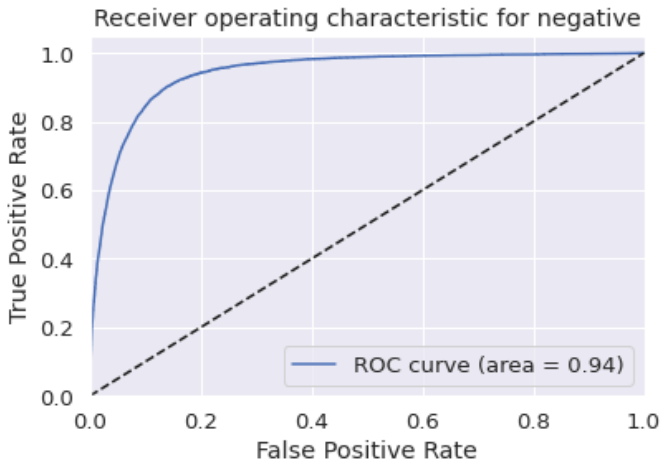

Figure 2: ROC curve for negative comments
Note : By construction, a false classification is in one of the two other categories. The AUC is in the label

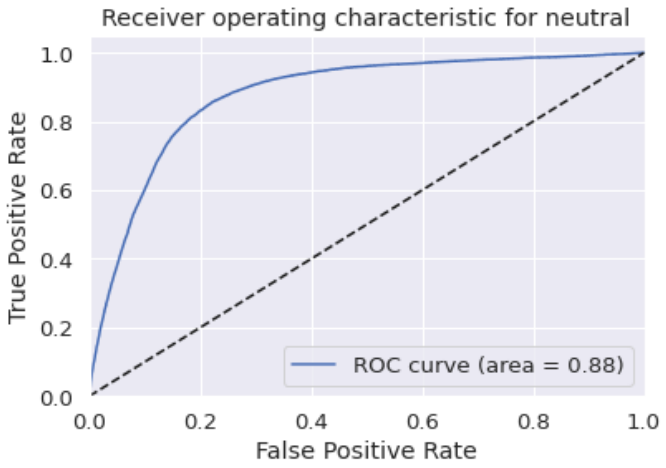

Figure 3: ROC curve for neutral comments
Note : By construction, a false classification is in one of the two other categories. The AUC is in the label

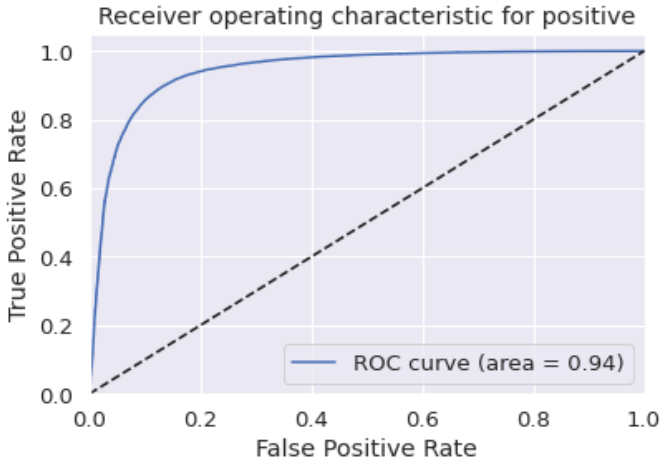

Figure 4: ROC curve for positive comments
Note : By construction, a false classification is in one of the two other categories. The AUC is in the label

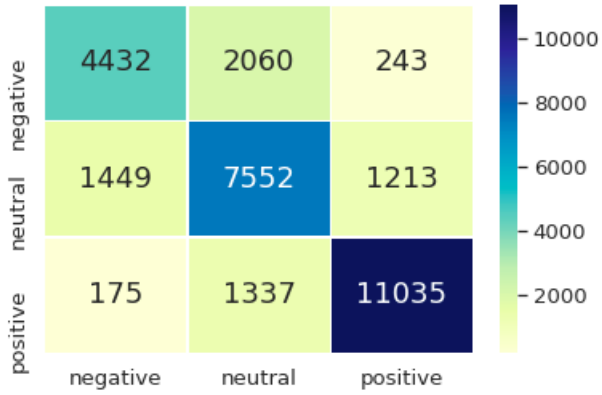

Figure 5: Confusion matrix of BERT
Note : x-axis is predicted labels. y-axis is true labels. Computed manually from the above matrix.

	Negative	Neutral	Positive
Negative	.66	.30	.04
Neutral	.14	.74	.12
Positive	.01	.11	.88

Table 3: Confusion matrix of BERT in percentage of true labels
Note : x-axis is predicted labels. y-axis is true labels. We find recall on the diagonal.

Label	Precision	Recall	F1-score
0	.73	.66	.69
1	.69	.74	.71
2	.88	.88	.88

Table 4: Classification report of BERT

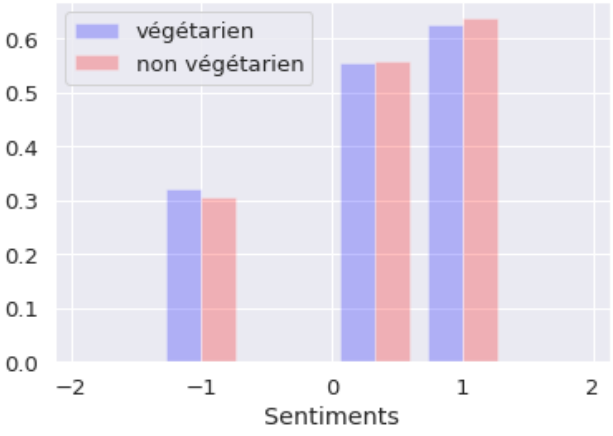

Figure 6: Histogram of the predicted sentiments for vegetarian and not vegetarian recipes.