Prediction of a Twitter user's location and hashtags
Machine Learning for Natural Language Processing 2022

Stanislas D'ORSETTI
3A parcours "Data Science, statistique apprentissage"
stanislas.dorsetti@ensae.fr

Eve PRAGER
MS Data Science
eve.prager@ensae.fr

Abstract

In this project, we choose to handle a dataset of tweets regarding the 2020 0lympic Games. The main goal of our study is to predict the location of a user according to the content of his tweet, his description and his indicated location. A complementary study will focus on the prediction of hashtags for a given tweet, among the 15 most commonly used. We rely on the content of the tweet. Finally, we implement two NLP models to answer a multi-class and multi-label classification problem, involving concepts presented in class.

1 Problem Framing

Twitter is a social network on which users are free to express their thoughts through tweets, small text of up to 280 characters. In this project, we analyze a database of 15,501 tweets concerning the 2020 Olympic Games, a world event that gather international athletes. Twitter's users widely commented these games, which took place in Tokyo in 2021. We asked ourselves two questions: where do the tweets about the games mainly come from, and what user are mentioning the most. In our dataset, a row corresponds to a tweet with mainly the author's user name, his self description and his location.

2 Experiments Protocol

We start with an exploratory phase on the data, using visualization. We note that on more than one out of five tweets the location or hashtag are not provided. We then adapt the database to our problem: we discard tweets not written in English as well as irrelevant variables. Furthermore, we remove the hashtags present in the tweet. We apply classic tweet tokenization methods to split tweets into words and punctuation. We also remove url links, hashtags, newlines and emojis. Thus we are able to visualize the most commonly used expressions using a multi-words expression detector.

2.1 Location

The location field is freely filled in by users, which makes its content very heterogeneous. In a majority of cases, the individual indicates either his city or his country or both. We also find aberrant names, emojis or any reference to the country of origin. We seek to find the origin of the tweet in a global way with classes by continents.

First, we apply the Name Entity Recognition on the variable 'user location' and we keep only the tweets that correspond to identified places. Then, we import a library which allows to obtain the latitude, the longitude and the identified country. Given this latitude and longitude, we make the labelization using shapefiles. The shapefile is a collection of point defining a polygon. We use a polygon of the continent that we consider to be labels, and if a point is in this polygon, we labellize with the associated continent. Given the geographic repartition of tweets, we focused on 4 labels : America, Europe, India and Oceania.

We tried different types of architecture for our model. First, we apply a word embedding with pretrained and fully connected vectors on it to transform the words into real vectors. Then, we apply the NER on the content of the tweet to keep only relevant part of the text. We trained the model both on the 'text' and 'user description' columns.

2.2 Hashtags

For this part, we discard tweets that do not contain hashtags. After a harmonization phase, we extract the most commonly used hashtags and plot descriptive statistics. Among most popular hash-

tags: Tokyo2020, Olympics and Mirabai Chanu[1].

We create the target variable containing a list of hashtags, by applying the following rule : if those are among the 15 most popular we keep them, otherwise we return an empty target. We then decompose as dummy variables, to place oneself within the framework of a multivariate binary classification problem. After stemming our data and add tokenization and embedding, we apply a "one vs rest" Logistic Regression for each hashtags.

3 Results

Finally, regarding the user location issue, we didn't manage to get good enough results. The best accuracy obtained was 0,6.

The problem of this approach is that it is not transferable to other datasets. There is a bias in the texts, because they all talk about the Olympic Games, and it would be difficult to transpose it to another dataset of tweets. Furthermore, there is a big issue related to the overrepresentation of India, as well so for the location that for the themes and subject of tweets. Even in the emojis, the most used, and by far, emoji is the indian flag...We tried to counterbalance this effect by using oversampling, but it did not seem to be sufficient enough. We were surprised that the user description and the NER method don't give better results.

Regarding the hashtags prediction, the results are however more encouraging. The logistic regression was able to predict almost all hashtags with an accuracy up to 0,95. We observe a negative correlation between the accuracy of the result and the popularity of the hashtag. Indeed, the more a hashtag is used, the worse predicted it is. The "Tokyo2020" hashtag is maybe the best example of that.

4 Discussion/Conclusion

For location prediction, maybe we could have also focused ourselves on fewer classes, and try to make a binary prediction, for example "India" against "Elsewhere". Our neural network model might be a bit too "naive" to be trained on these complex data and to learn the relationship between the text and the localisation, that is not so easy to catch, even for humans with wordclouds... With user description,it would be interesting to restrain the number of tokens associated with NER and to transform the input data in dummies variables with, for example, Marabai Channu, India and Sweden as features, based on the wordclouds that showed us that these words were particularly discriminant regarding the location of the writer. Finally, we could have also concatenated these models and make a final prediction issued from a dense layer at the very end of the network. We though about other models, such as CNN, or putting fewer possible inputs for the NER approach, based on the wordclouds.

For hashtag prediction, we could have used the Pytorch deep learning library, trained it with BERT data and then implement models simply with pytorch lightning.

Other ideas to extend our research would be to use emojis to predict the location or the hashtags of a tweet.

Finally, we find this project very interesting as it relies on real data and allow us to apply NLP methods.

Please find our github at this adress : https://github.com/sdorsetti/NLP_project Please find our GoogleColab at this adress : https://colab.research.google.com/drive/1EnKuPLVvHGfaObWxqXI6E4XYkcK6FQXn#scrollTo=dg0O58OkEv0U

.

[1] indian weightlifter that won silver medal on these Games

References

[1] Rahmanzadeh Heravi, B. and Salawdeh, I., "Tweet Location Detection" *http://cj2015.brown.columbia.edu/papers/tweet-location.pdf*, 2015.

[2] Kartik Nooney, "Deep dive into multi-label classification..! (With detailed Case Study)", *https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff*, Towards Data Science, 2018.

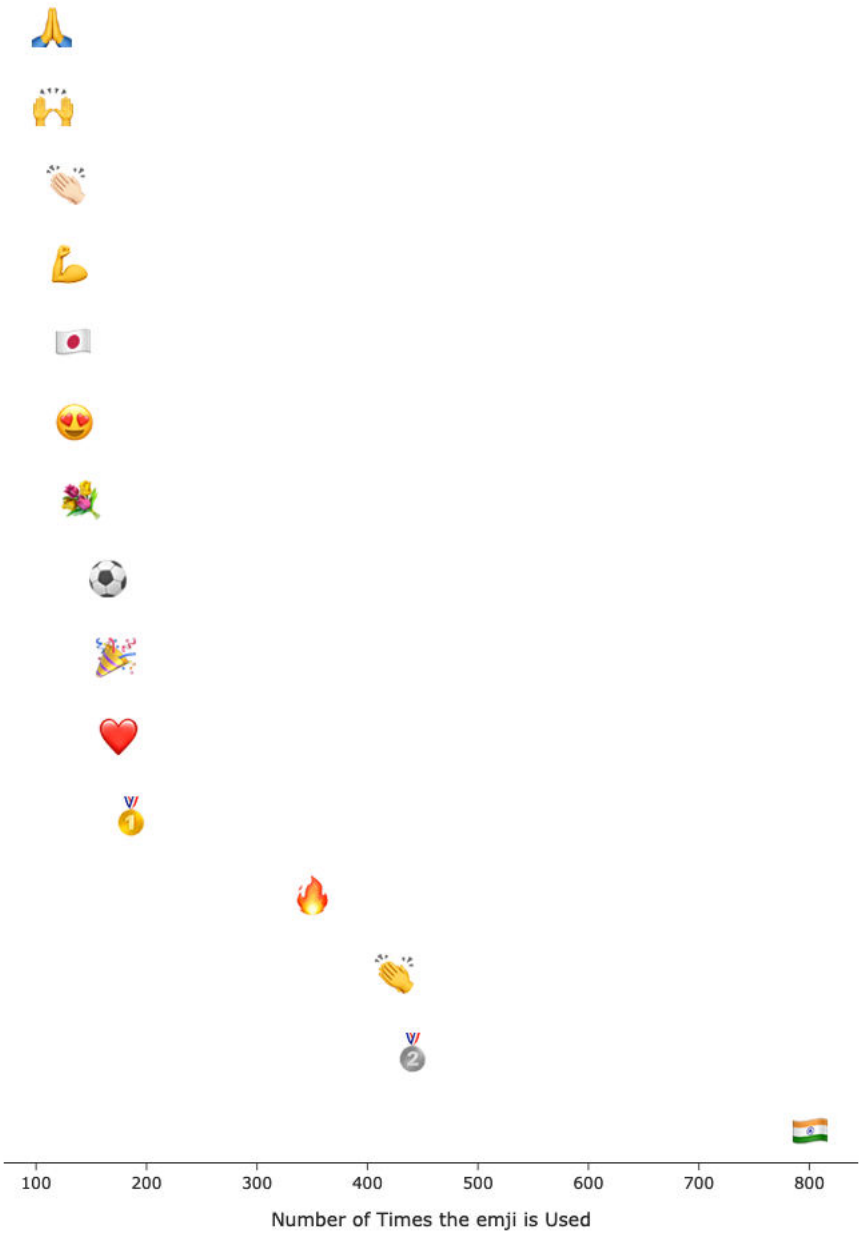


Figure 1: Emoji's occurence in the dataset

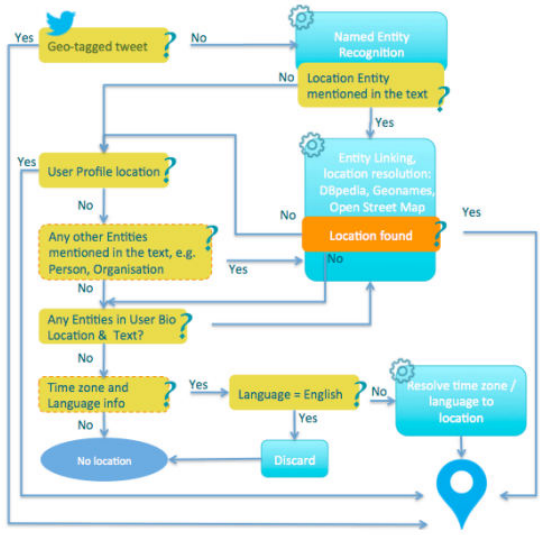


Figure 1. Twiloc: A Tweet Location Detection Framework

Figure 2: Location prediction model, from the litterature [1]

Figure 3: Localisation of random tweets

Figure 4: ”Real”repartition of localisations with the considered labels

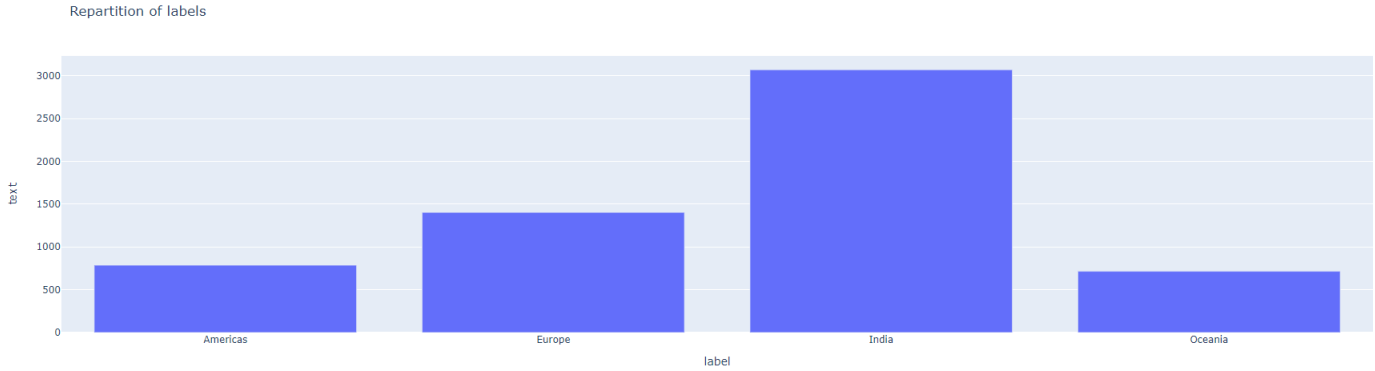


Figure 5: Validation and Training loss, on epochs for the model during the neural network training for user location prediction

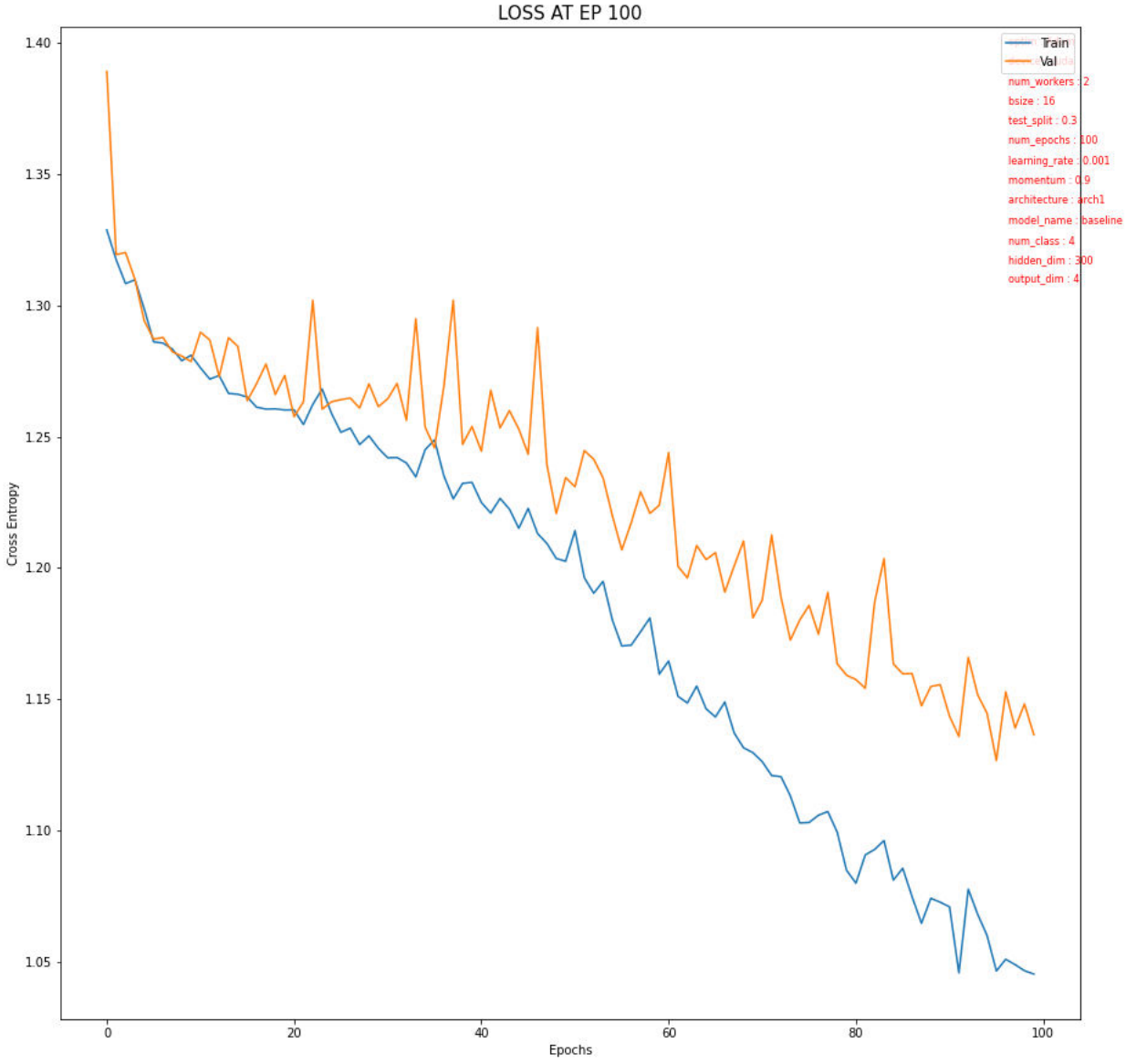
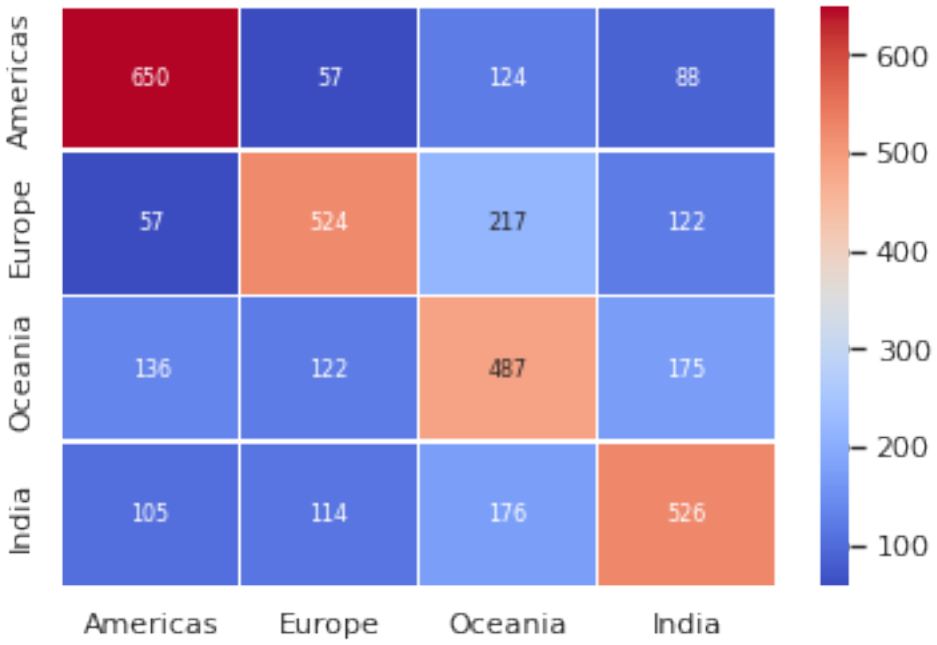


Figure 6: Confusion matrix of results for the user locations prediction for the best model



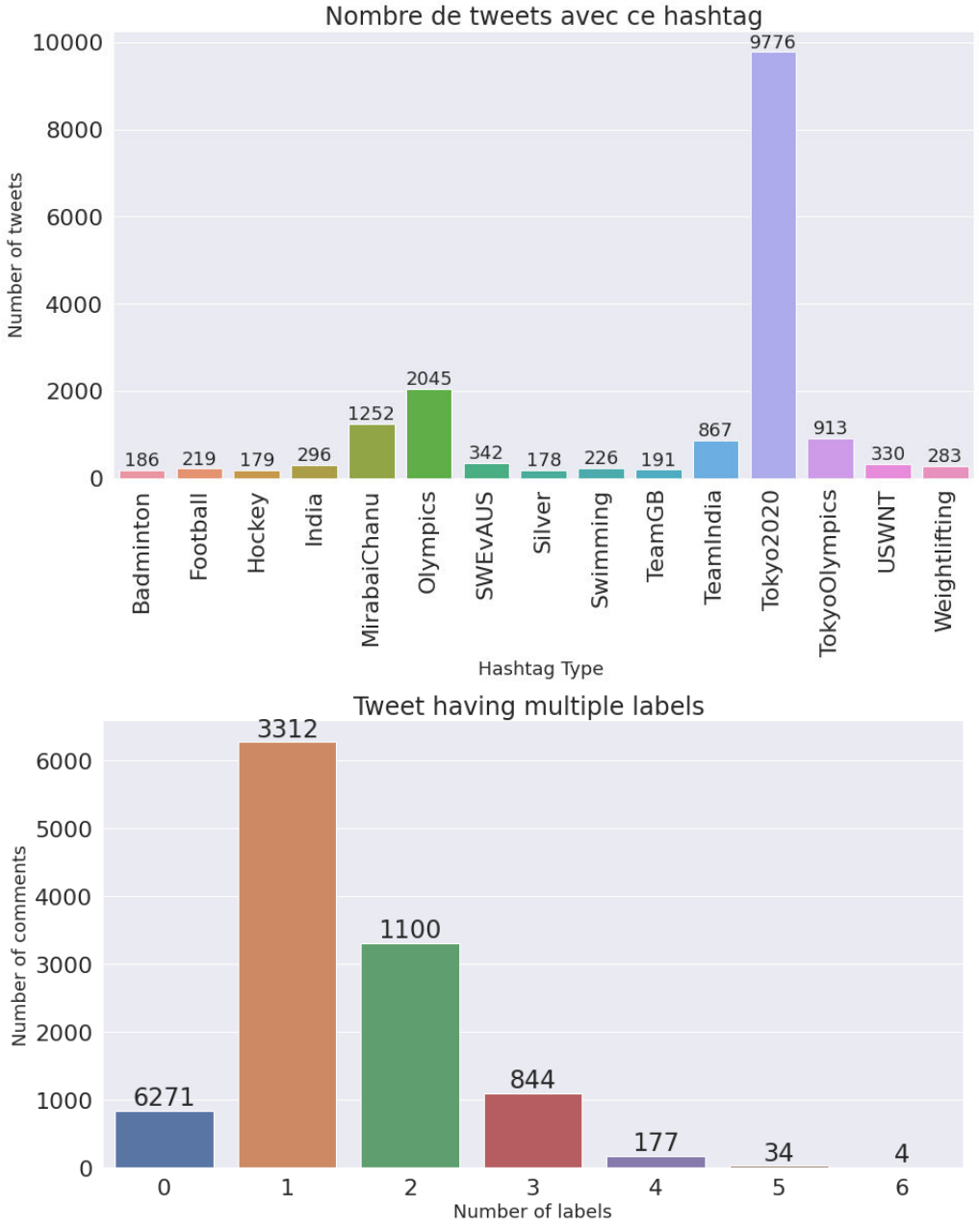


Figure 7: Occurence of most common hashtags

Figure 8: Confusion matrix of results for Logistic Regression one vs rest, for worse predicted hashtag