Identifying Brazilian Names using a Recurrent Neural Network

By Duaa Tashkandi, Wjdan Alharthi, Ben Gaudiosi, and Shreya Ramesh

Introduction

In an effort to help Digaai, an online platform that aggregates and tracks the Brazilian diaspora and its cultural influence, the goal of our project is to classify a person's full name as either Brazilian or not Brazilian. We came up with several different models, including a logistic regression, a vanilla neural network, and a recurrent neural network. Using these machine learning techniques, we've managed to develop a fairly accurate model for determining whether or not a person is Brazilian based on their first and last name.

Datasets

For Brazilian name sets, we scraped Facebook pages of Brazilian groups, for example, "Clash of Clans Brazil" or "Brazilians for Animal Rights." For non-Brazilian data, we scraped Facebook pages of colleges to ensure ethnic diversity in the name set of non-Brazilians. The final Distribution of Brazilian and non-Brazilian classes is 1:1. We have collected a total of 60,000 names. Therefore, we have a baseline accuracy of 50% we aimed to beat, which would be random guessing.

We used 80% of our data for training and 20% for testing and validation.

RNN Model

A recurrent neural network is different from a vanilla one because it learns from sequences, and each input has an associated timestamp. In our implementation, each letter of the name would be an input to the network with the position as the timestamp, e.g. the first letter of the name would have a timestamp of 1.

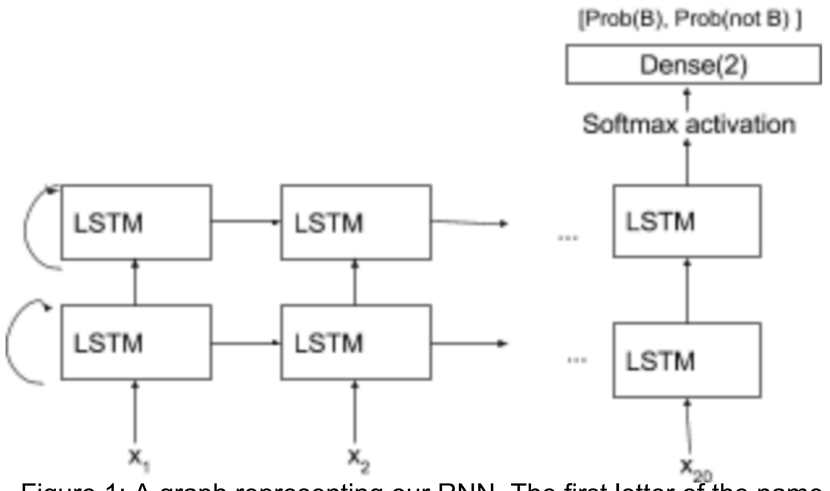

Figure 1: A graph representing our RNN. The first letter of the name is X_1, the second X_2, etc. "LSTM" stands for Long-short term memory which is a block used for building connections between inputs.

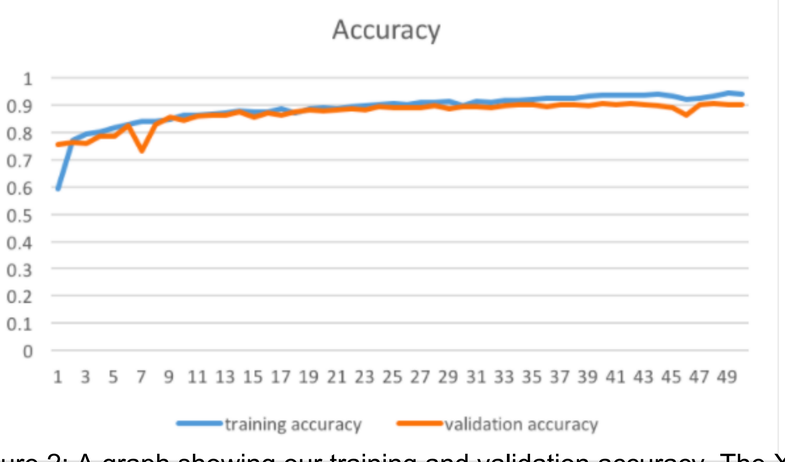

Figure 2: A graph showing our training and validation accuracy. The X axis is number of epochs, while the Y axis is the accuracy of the model.

Evaluation

We first tried the RNN on just first name inputs, and got an accuracy of 86% on testing data. We then trained the model on inputs of first name and last name as two separate features and got an accuracy of 90% on the testing data. Figure 2 compares the training accuracy to the validation accuracy.

Conclusion

In the end, the RNN was the most successful because it took into account the order of the characters. There are several ways we imagine the RNN could be improved. First, we could use pre-trained character embedding instead of one-hot encoding. Another improvement would be better and leaner sampling of data, from a more accurate source than Facebook. We hope that the model we've created here will be useful in achieving Digaai's goal of tracking the Brazilian cultural diaspora.

References

- Treeratpituk, Pucktada, and C. Lee Giles. "Name-Ethnicity Classification and Ethnicity-Sensitive Name Matching." AAAI. 2012
- Lee, Kim, Ko, D. Choi, J. Choi, Kang. "Name Nationality Classification with Recurrent Neural Networks." IJCAI. 2017.