Bag of Authors

Leonor Furtado, Pedro Carvalho, Carolina Simão
M20190308, M20190417, M20190418

Text Mining

1 Introduction

In this report, a solution is presented for the Bag-of-authors problem, which was to predict which of the 6 authors wrote each of the 500- and 1000-word text excerpts. The objective is to create a model that receives as input the excerpts provided for training and learns from that provided labeled set to identify the authors given in the test folder.

2 Method/Approach

2.1 Preprocessing

Different approaches were used in the preprocessing phase, such as removing stop words, lemmatization or stemming, resampling the data into either 500 or a 1000-word chunks, and also different types of vectors: bag-of-words and TF-IDF.

In the end, some manual cleansing of the Almada Negreiros and Luisa Marques Silva was also performed as there was a lot of noise with editorial information that wasn't present in the samples we were trying to predict. This was done in an effort to align our train scenario with our prediction problem scenario.

In order to improve on the baseline version of the project several strategies were attempted:

1. Breaking down the training samples into chunks of:
    a. 500 words
    b. 1000 words

2. Reduce morphological variation in the text:
    a. Lemmatisation
    b. Stemming

3. Vectorisation:
    a. Bag-of-words (using different n-grams)
    b. TF-IDF

2.2 Models

Different models were tested with increasing complexity in order to obtain better accuracy scores.

In order to attempt to improve model performance several train/dev/test strategies were attempted, to try and solve the problem of a highly unbalanced dataset, with Luisa and Almada having much lower train samples than the other authors, the following were attempted across the different models:

1. Having a hold-out set of a full title for each author for testing, to avoid overfitting
2. Random sampling of the samples at a 80/20 ratio of the train files for train/test set respectively
3. Under-sampling over all the authors to correct the imbalance in the dataset

2.2.1 K-Nearest Neighbors

The project was initially developed with a baseline model based in K-Nearest Neighbors

For KNN, the data was split in 80/20% for training and testing sets respectively. Lemmatized data and the TF-IDF vectorization was used. For the KNN classifier, 5 neighbors, cosine metric and a leaf size of 30 was chosen. Many variations of these hyperparameters were attempted including the different combinations

of the preprocessing steps described in section 2.1. The best configuration that yielded the best results was 1. a) , 2.a) and 3.b) together with under-sampling over all the authors to correct the imbalance of the train dataset.

2.2.2 Naïve Bayes (NB)

Since one of titles of one of the authors was still missing using the KNN model, the Naive Bayes approach was also implemented.

For NB, the data was also split in 80/20% for training and testing sets respectively. Again, lemmatized data and the TF-IDF vectorization were used. As with KNN Many variations of these hyperparameters were attempted including the different combinations of the preprocessing steps described in section 2.1. The best configuration that yielded the best results was 1. a), 2.a) and 3.b) together with under-sampling over all the authors to correct the imbalance of the train dataset.

2.2.3 Long Short-Term Memory (LSTM)

Finally, a LSTM model was applied to the data, in order to try to predict the whole set of authors that the other two methods failed to do.

Only the basic removal of punctuation, html tags and lowercasing were performed for preprocessing was used except for the chunking of the data.

And the train/test split strategy was that of the hold-out of an entire title for each author described in 2.2 item 1, which was used as a validation set when training the LSTM model.

The design of the model consists of an initial embedding layer that reduces the input size from the maximum number of words parameter in our tokenizer to an output dimension of 100. A 20% dropout layer was put in between these to avoid overfitting, especially given our unbalanced dataset. Followed by an LSTM layer with 100 neurons are added to the model (also having 20% dropout), followed by a softmax output layer with size 6 (the number of possible labels for our classifier). We then take the maximum probability label to be our final predictor.

To improve our model other model configurations were attempted but training proved to be too slow to yield any significant results. This included adding a bidirectional component to the LSTM layer, as well as an extra dense layer in-between the LSTM layer and the output layer with relu activation.

High accuracies were achieved, but there was also very low precision in some authors, implying either some overfitting or that the data is still unbalanced. The randomness involved in neural networks yielded better results for Almada at one time and Luisa at other times.

3 Results and Discussion

3.1 Results

3.1.1 K-Nearest Neighbors

With the KNN approach the best results were obtained. The accuracy for the test set was 94% but when applied to the unlabeled dataset, it only returned an 100% accuracy, with 100% recall for the author Luisa Marques Silva.

AlmadaNegreiros	0.96	0.96	0.96	28
CamiloCasteloBranco	0.84	0.94	0.89	17
EcaDeQueiros	1.00	0.93	0.96	14
JoseRodriguesSantos	1.00	0.94	0.97	16
JoseSaramago	1.00	0.87	0.93	15
LuisaMarquesSilva	0.90	1.00	0.95	18
accuracy			0.94	108
macro avg	0.95	0.94	0.94	108
weighted avg	0.95	0.94	0.94	108

	precision	recall	f1-score	support
AlmadaNegreiros	1.00	1.00	1.00	2
CamiloCasteloBranco	1.00	1.00	1.00	2
EcaDeQueiros	1.00	1.00	1.00	2
JoseRodriguesSantos	1.00	1.00	1.00	2
JoseSaramago	1.00	1.00	1.00	2
LuisaMarquesSilva	1.00	1.00	1.00	2
accuracy			1.00	12
macro avg	1.00	1.00	1.00	12

3.1.2 Naïve Bayes

The accuracy for the test set was 95% but when applied to the unlabeled dataset, it only returned an 83% accuracy, with 0% recall for the author Luisa Marques Silva. The Naive Bayes model failled to predict Luisa Marques Silva.

	precision	recall	f1-score	support
AlmadaNegreiros	1.00	0.94	0.97	18
CamiloCasteloBranco	0.90	1.00	0.95	18
EcaDeQueiros	1.00	0.94	0.97	18
JoseRodriguesSantos	0.95	1.00	0.97	18
JoseSaramago	0.94	0.94	0.94	18
LuisaMarquesSilva	0.94	0.89	0.91	18
accuracy			0.95	108
macro avg	0.96	0.95	0.95	108

	precision	recall	f1-score	support
AlmadaNegreiros	1.00	1.00	1.00	2
CamiloCasteloBranco	1.00	1.00	1.00	2
EcaDeQueiros	1.00	1.00	1.00	2
JoseRodriguesSantos	0.50	1.00	0.67	2
JoseSaramago	1.00	1.00	1.00	2
LuisaMarquesSilva	0.00	0.00	0.00	2
accuracy			0.83	12
macro avg	0.75	0.83	0.78	12
weighted avg	0.75	0.83	0.78	12

Label predictions for both KNN and NB are presented in the table below with column names pred_KNN and pred_NB, respectively:

number_of_words	y_true	pred_KNN	pred_NB
1000Palavras/text1.txt	JoseSaramago	JoseSaramago	JoseSaramago
1000Palavras/text2.txt	AlmadaNegreiros	AlmadaNegreiros	AlmadaNegreiros
1000Palavras/text3.txt	LuisaMarquesSilva	LuisaMarquesSilva	JoseSaramago
1000Palavras/text4.txt	EcaDeQueiros	EcaDeQueiros	EcaDeQueiros
1000Palavras/text5.txt	CamiloCasteloBranco	CamiloCasteloBranco	CamiloCasteloBranco
1000Palavras/text6.txt	JoseRodriguesSantos	JoseRodriguesSantos	JoseRodriguesSantos
500Palavras/text1.txt	JoseSaramago	JoseSaramago	JoseSaramago
500Palavras/text2.txt	AlmadaNegreiros	AlmadaNegreiros	AlmadaNegreiros
500Palavras/text3.txt	LuisaMarquesSilva	LuisaMarquesSilva	JoseSaramago
500Palavras/text4.txt	EcaDeQueiros	EcaDeQueiros	EcaDeQueiros
500Palavras/text5.txt	CamiloCasteloBranco	CamiloCasteloBranco	CamiloCasteloBranco
500Palavras/text6.txt	JoseRodriguesSantos	JoseRodriguesSantos	JoseRodriguesSantos

3.1.3 Long Short-Term Memory

With the Long Short-Term Memory Neural Network the maximum accomplished accuracy was 75%, with only 50 neurons in the LSTM layer and else as described in section 2.2.3.

	precision	recall	f1-score	support
AlmadaNegreiros	0.00	0.00	0.00	2
CamiloCasteloBranco	1.00	1.00	1.00	2
EcaDeQueiros	0.50	1.00	0.67	2
JoseRodriguesSantos	1.00	1.00	1.00	2
JoseSaramago	0.67	1.00	0.80	2
LuisaMarquesSilva	1.00	0.50	0.67	2
accuracy			0.75	12
macro avg	0.69	0.75	0.69	12
weighted avg	0.69	0.75	0.69	12

The usual 4 authors with the most data were predicted correctly, even though we lost the power to predict Almada, the author Luisa Marques Silva also showed up.

number_of_words	y_true	predicted
1000Palavras/text1.txt	JoseSaramago	JoseSaramago
1000Palavras/text2.txt	AlmadaNegreiros	EcaDeQueiros
1000Palavras/text3.txt	LuisaMarquesSilva	JoseSaramago
1000Palavras/text4.txt	EcaDeQueiros	EcaDeQueiros
1000Palavras/text5.txt	CamiloCasteloBranco	CamiloCasteloBranco
1000Palavras/text6.txt	JoseRodriguesSantos	JoseRodriguesSantos
500Palavras/text1.txt	JoseSaramago	JoseSaramago
500Palavras/text2.txt	AlmadaNegreiros	EcaDeQueiros
500Palavras/text3.txt	LuisaMarquesSilva	LuisaMarquesSilva
500Palavras/text4.txt	EcaDeQueiros	EcaDeQueiros
500Palavras/text5.txt	CamiloCasteloBranco	CamiloCasteloBranco
500Palavras/text6.txt	JoseRodriguesSantos	JoseRodriguesSantos

4 Conclusion

To conclude, it was quite hard to provide satisfactory results from such an imbalanced dataset and it required multiple instances of testing with various methods of processing and different models as well. We overcame this challenge successfully by performing under-sampling to correct our class imbalance.

This couldn't be applied to all our models, since for example the deep learning LSTM model required the full power of the large dataset to yield satisfactory results, tuning the model proved difficult given the high training time and the number of hyperparameters.

It has also been shown that the most complicated model is not, necessarily, the best model, which was in this case the lemmatized vectors in KNN. There is still some room for improvement since the model's sensitivity to the less represented classes could be higher but it does predict well most labels.

This project was a fine introduction to the topic of NLP, that required a broad usage of different techniques and creativity to solve the proposed problem. In this sense fulfilled its purpose and this group feels less unprepared to tackle real world problems when the time arises.

5 Appendix

When using the original unchanged imbalanced dataset split in 500-word samples and the KNN model using the same parameters as described in section 2.2.1. The accuracy for the test set was 97% but when applied to the unlabeled dataset, it only returned an 83% accuracy, with 0% recall for the author Luisa Marques Silva.

	precision	recall	f1-score	support
AlmadaNegreiros	1.00	0.64	0.78	25
CamiloCasteloBranco	0.97	0.98	0.97	329
EcaDeQueiros	0.97	0.92	0.95	179
JoseRodriguesSantos	0.96	0.99	0.98	447
JoseSaramago	0.97	0.98	0.98	400
LuisaMarquesSilva	0.87	0.76	0.81	17
accuracy			0.97	1397
macro avg	0.96	0.88	0.91	1397
weighted avg	0.97	0.97	0.97	1397

	precision	recall	f1-score	support
AlmadaNegreiros	1.00	1.00	1.00	2
CamiloCasteloBranco	1.00	1.00	1.00	2
EcaDeQueiros	1.00	1.00	1.00	2
JoseRodriguesSantos	0.50	1.00	0.67	2
JoseSaramago	1.00	1.00	1.00	2
LuisaMarquesSilva	0.00	0.00	0.00	2
accuracy			0.83	12
macro avg	0.75	0.83	0.78	12
weighted avg	0.75	0.83	0.78	12

Five out of the six authors were sucessfully predicted, with an accuracy of 83%.

number_of_words	y_true	text	pred_KNN
1000Palavras/text1.txt	JoseSaramago	pouco pouco tranquilidade regressa agora lidi…	JoseSaramago
1000Palavras/text2.txt	AlmadaNegreiros	justamente tido ideia fazer cabeça christo christo int…	AlmadaNegreiros
1000Palavras/text3.txt	LuisaMarquesSilva	quase mês época exames aproximava-se vertiginosamente ac…	JoseRodriguesSantos
1000Palavras/text4.txt	EcaDeQueiros	agora porém fervor arrastadamente elle levava bosque onde …	EcaDeQueiros
1000Palavras/text5.txt	CamiloCasteloBranco	cahos cima descer descer mortalha treva sobre abysmo …	CamiloCasteloBranco
1000Palavras/text6.txt	JoseRodriguesSantos	senhor ensina pena homem sabe alá fala directamente crente…	JoseRodriguesSantos
500Palavras/text1.txt	JoseSaramago	pouco pouco tranquilidade regressa agora lidi…	JoseSaramago
500Palavras/text2.txt	AlmadaNegreiros	justamente tido ideia fazer cabeça christo christo int…	AlmadaNegreiros
500Palavras/text3.txt	LuisaMarquesSilva	quase mês época exames aproximava-se vertiginosamente ac…	JoseRodriguesSantos
500Palavras/text4.txt	EcaDeQueiros	agora porém fervor arrastadamente elle levava bosque onde …	EcaDeQueiros
500Palavras/text5.txt	CamiloCasteloBranco	cahos cima descer descer mortalha treva sobre abysmo …	CamiloCasteloBranco
500Palavras/text6.txt	JoseRodriguesSantos	senhor ensina pena homem sabe alá fala directamente crente…	JoseRodriguesSantos

When using the original imbalanced dataset split in 500-word samples and the NB model using the same parameters as described in section 2.2.2.

Even though an 95% accuracy for the test set was obtained, the unlabeled dataset only returned an unsatisfactory 67% accuracy, with 0% precision not just for the author Luisa but also for Almada Negreiros.

	precision	recall	f1-score	support
AlmadaNegreiros	0.00	0.00	0.00	20
CamiloCasteloBranco	0.94	0.99	0.96	310
EcaDeQueiros	0.96	0.92	0.94	179
JoseRodriguesSantos	0.95	1.00	0.97	463
JoseSaramago	0.97	0.97	0.97	407
LuisaMarquesSilva	0.00	0.00	0.00	18
accuracy			0.95	1397
macro avg	0.63	0.65	0.64	1397
weighted avg	0.93	0.95	0.94	1397

	precision	recall	f1-score	support
AlmadaNegreiros	0.00	0.00	0.00	2
CamiloCasteloBranco	1.00	1.00	1.00	2
EcaDeQueiros	0.50	1.00	0.67	2
JoseRodriguesSantos	0.50	1.00	0.67	2
JoseSaramago	1.00	1.00	1.00	2
LuisaMarquesSilva	0.00	0.00	0.00	2
accuracy			0.67	12
macro avg	0.50	0.67	0.56	12
weighted avg	0.50	0.67	0.56	12

With this instance of the NB model only four out of five authors were successfully predicted.

number_of_words	y_true	text	pred_KNN	pred_NB
1000Palavras/text1.txt	JoseSaramago	pouco pouco tranquilidade regressa agora lidi…	JoseSaramago	JoseSaramago
1000Palavras/text2.txt	AlmadaNegreiros	justamente tido ideia fazer cabeça christo christo int…	AlmadaNegreiros	EcaDeQueiros
1000Palavras/text3.txt	LuisaMarquesSilva	quase mês época exames aproximava-se vertiginosamente ac…	JoseRodriguesSantos	JoseRodriguesSantos
1000Palavras/text4.txt	EcaDeQueiros	agora porém fervor arrastadamente elle levava bosque onde …	EcaDeQueiros	EcaDeQueiros
1000Palavras/text5.txt	CamiloCasteloBranco	cahos cima descer descer mortalha treva sobre abysmo …	CamiloCasteloBranco	CamiloCasteloBranco
1000Palavras/text6.txt	JoseRodriguesSantos	senhor ensina pena homem sabe alá fala directamente crente…	JoseRodriguesSantos	JoseRodriguesSantos
500Palavras/text1.txt	JoseSaramago	pouco pouco tranquilidade regressa agora lidi…	JoseSaramago	JoseSaramago
500Palavras/text2.txt	AlmadaNegreiros	justamente tido ideia fazer cabeça christo christo int…	AlmadaNegreiros	EcaDeQueiros
500Palavras/text3.txt	LuisaMarquesSilva	quase mês época exames aproximava-se vertiginosamente ac…	JoseRodriguesSantos	JoseRodriguesSantos
500Palavras/text4.txt	EcaDeQueiros	agora porém fervor arrastadamente elle levava bosque onde …	EcaDeQueiros	EcaDeQueiros
500Palavras/text5.txt	CamiloCasteloBranco	cahos cima descer descer mortalha treva sobre abysmo …	CamiloCasteloBranco	CamiloCasteloBranco
500Palavras/text6.txt	JoseRodriguesSantos	senhor ensina pena homem sabe alá fala directamente crente…	JoseRodriguesSantos	JoseRodriguesSantos

Since the results were worse than KNN, the Naive Bayes approach was deemed unhelpful. However, when the under-sampled dataset was used better results were yielded for both KNN and NB, that is why these results are now being presented in the appendix while the new improved ones are presented in the main body of the report.

P.S: If you would like to be added to the GitHub private repository used for the development of this report please email your GitHub username to:

M20190308@novaims.unl.pt