

Spooky Author Identification

1. Problem description

The dataset contains text from works of fiction written by spooky authors of the public domain. The data was prepared by chunking larger texts into sentences using CoreNLP's MaxEnt sentence tokenizer

This is a multi-labels classification, with 3 labels:

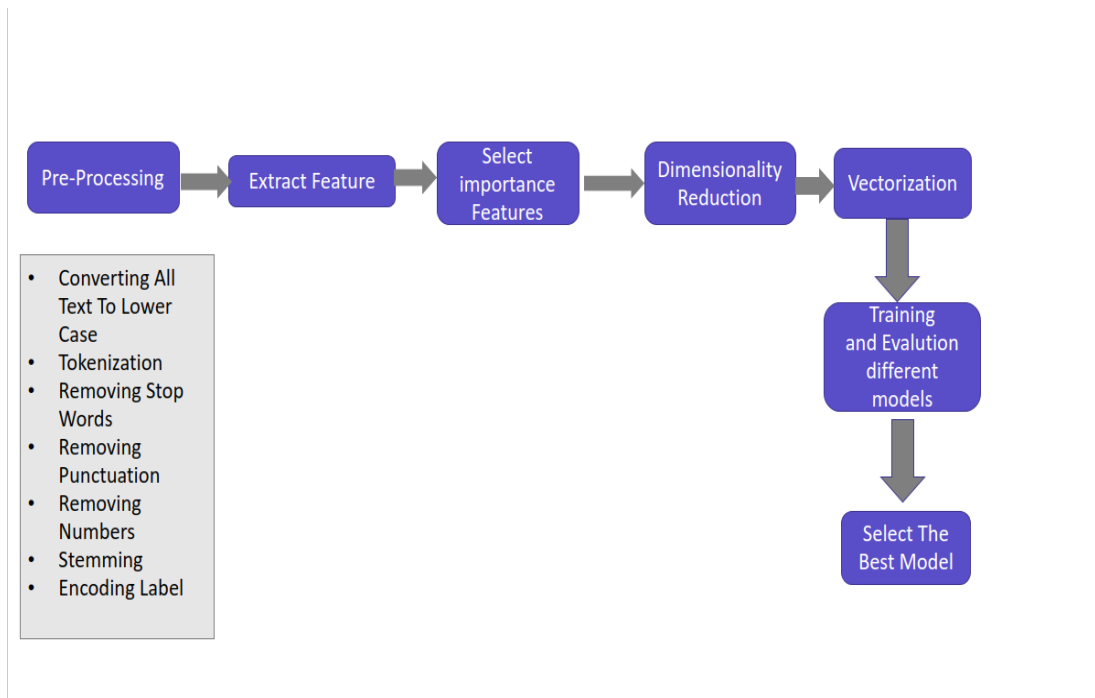
EAP: short text written by Edgar Allan Poe

MWS: short text written by Mary Shelley

HPL: short text written by HP Lovecraft

Goal : we need to build a model which is able to know the author depend on the phrase(author's writing style)

2. Model design



3. Experimental results

We try different Models to achieve the best accuracy :

The Result with Naïve Base Model on test data :

	precision	recall	f1-score	support
0	0.83	0.84	0.84	1570
1	0.85	0.84	0.84	1071
2	0.85	0.85	0.85	1275
accuracy			0.84	3916
macro avg	0.84	0.84	0.84	3916
weighted avg	0.84	0.84	0.84	3916

The Result with SGD Classifier Model on test data:

	precision	recall	f1-score	support
0	0.85	0.79	0.82	1570
1	0.82	0.79	0.80	1071
2	0.78	0.88	0.82	1275
accuracy			0.82	3916
macro avg	0.82	0.82	0.81	3916
weighted avg	0.82	0.82	0.82	3916

4. Model performance:

Because of Naïve Base Model is more accurate we will depend on it

The Confusion Matrix of it is :

	precision	recall	f1-score	support
0	0.83	0.84	0.84	1570
1	0.85	0.84	0.84	1071
2	0.85	0.85	0.85	1275
accuracy			0.84	3916
macro avg	0.84	0.84	0.84	3916
weighted avg	0.84	0.84	0.84	3916