Natural Language Mini-project 2

Group 24, Diogo Moura 86976, Catarina Machuqueiro 89423

1 EXPLORATORY DATA ANALYSIS

The first step in finding a solution for a problem is to first analyse the available data. We found that the most frequent terms in the dataset were punctuation, stop words, and a few numbers. We also found that the classes were imbalanced.

Figure 1: Word Cloud without text preprocessing

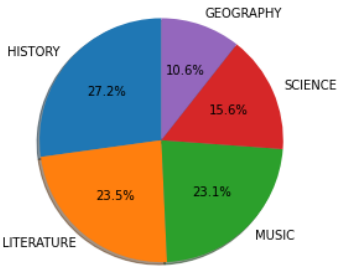

Figure 2: Labels Distribution

2 MODELS

The solution uses a Voting Classifier that has as estimators a Ridge classifier (**Ridge**), an Extra-trees classifier (**Extra Trees**), a C-Support Vector classifier (**SVC**), and a Complement Naive Bayes classifier (**Complement NB**).

This way, we can get a good robust score independently of the data which will most likely yield better results, independently of the unseen data and reduce the number of situations where our system performs poorly. As a downside, the classification will take more time, due to the computation of several scores/probabilities. We used TF-IDF, which gives less importance to more common words.

In the preprocessing phase, we make the following changes:

(1) Remove punctuation;
(2) Remove single letters;
(3) Replace roman numerals with the keyword "roman_numeral";
(4) Replace ordinal numbers with "ordinal_number";
(5) Replace 4-digit numbers with "year";
(6) Replace the remaining numbers with "number";
(7) Replace words that correspond to months with "month";
(8) Reduce the words to their word stem using the Snowball Stemmer;

The word distribution changes drastically with preprocessing.

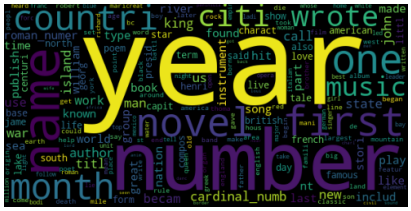

Figure 3: Word Cloud with text preprocessing

The removal of punctuation and the replacement of certain words by keywords that represent them accounts for benefits in performance in the classification system and memory occupied by the data, without loss of information.

3 EXPERIMENTAL SETUP

In order to discover the best solution for the problem, we conducted several experiments, including different forms of preprocessing, different vectorization algorithms, different classifiers, and model parameterization. The combination that yielded the best results was kept.

To evaluate the model we used the following metrics:

- Accuracy: calculates the fraction of predictions our model got right;
- Balanced Accuracy: computes the average of recall obtained on each class;
- F1-Score: aggregated measure of precision and recall.

Since the dataset is not very large, we used 5-fold cross-validation with a random-shuffle of the data to evaluate the model. This validation allowed us to increase the confidence in our experimental setup training and evaluation.

We used as a baseline a model based on a Support Vector Machine (SVM) without any preprocessing and using count vectorizer.

4 RESULTS

Our baseline model has an accuracy of 81.3%, as can be seen on table 1.

First, we compared two vectorization algorithms: one that converts the text to a matrix of token counts (**Count Vectorizer**) and another to a matrix of TF-IDF features (**TF-IDF Vectorizer**).

Table 1: Results with Cross-Validation without preprocessing of the SVC classifier using different vectorization algorithms

Vectorizer	Acc.	Bal. Acc.	F1-Score
Count	0.813	0.808	0.811
TF-IDF	0.833	0.819	0.83

TF-IDF is the best vectorizer. This can be explained by the fact that, unlike Count Vectorizer, TF-IDF doesn't give too much importance to values that show up in the documents in large quantities (e.g. stopwords).

Afterwards, we experimented with different classifiers.

Table 2: Results with Cross-Validation without preprocessing

Classifier	Acc.	Bal. Acc.	F1-Score
Ridge	0.849	0.838	0.846
SVC	0.833	0.819	0.83
Complement NB	0.855	0.846	0.849
Extra Trees	0.755	0.731	0.749
Voting	0.847	0.835	0.844

All the models produce similar results, although the voting classifier is the most robust, as previously stated.

The use of POS tagging as a feature was experimented. Each term instead of being represented as its stem was instead represented as the concatenation of its stem with its POS tag. This approach did not increase the accuracy of our model and drastically increased the dimensionality of the problem, since each term was now treated differently based on its POS tag.

Then, we tried different types of preprocessing.

Table 3: Results with Cross-Validation and preprocessing with stopword removal

Classifier	Acc.	Bal. Acc.	F1-Score
Ridge	0.846	0.838	0.843
SVC	0.843	0.834	0.842
Complement NB	0.849	0.847	0.844
Extra Trees	0.792	0.784	0.79
Voting	0.85	0.844	0.848

Table 4: Results with Cross-Validation and preprocessing without stop-words removal

Classifier	Acc.	Bal. Acc.	F1-Score
Ridge	0.85	0.841	0.847
SVC	0.842	0.833	0.84
Complement NB	0.854	0.85	0.849
Extra Trees	0.791	0.777	0.788
Voting	0.853	0.845	0.852

The stop-words are not very informative. However, removing them does not improve our classification model, since we use measures that compensate for this, such as TF-IDF.

As can be seen, the balanced accuracy is fairly good and for this reason we did not use anything to reduce the impact of class imbalance in our classification, also because the class imbalance is very small.

Table 5: Final Results without cross-validation

Classifier	Acc.	Bal. Acc.	F1-Score
Ridge	0.898	0.887	0.894
SVC	0.878	0.868	0.879
Complement NB	0.88	0.88	0.881
Extra Trees	0.824	0.809	0.821
Voting	0.900	0.894	0.899

We consider our best model as the voting classifier, so our final accuracy in the test dataset is 90%.

5 ERROR ANALYSIS

The overlapping between topics can be visualized in table 6:

Table 6: Overlap in terms between topics

	History	Geography	Science	Music	Literature
History	6802	1515	1953	2331	2581
Geography	-	2814	974	1057	1151
Science	-	-	4675	1655	1746
Music	-	-	-	5769	2322
Literature	-	-	-	-	6015

The topics with more overlapping terms will be, as expected, more difficult to distinguish.

As can be seen on table 7, the categories most difficult to distinguish are History and Music, which have the biggest number of overlapping terms!

Table 7: Confusion Matrix of the Voting Classifier classification

True \ Predicted	History	Geography	Science	Music	Literature
History	0.891	0.029	0.022	0.022	0.036
Geography	0.15	0.85	0	0	0
Science	0.091	0	0.886	0.011	0.011
Music	0.064	0.009	0.009	0.891	0.027
Literature	0.056	0	0.008	0.008	0.927

Analysing the confusion matrix, Geography is the category with the lowest accuracy and Literature has the highest accuracy.

6 FUTURE WORK

The use of Neural Networks as classifiers could yield better results in the classification. Additionally, to reduce the problem dimensionality without loss of information, we could use word embeddings using, for example, the Glove model. Furthermore, we could use sentence embeddings. The use of embeddings would increase the classification efficiency and possibly even yield better results.