Intelligent-Systems:
NLP on Research Articles

Author: Julia Sánchez Martínez

January 11, 2022

1 Introduction

Natural Language Processing (NLP) is understood as the study of the interactions between computers and the human natural language, such as speech or text. Working with NL is more challenging than working with other types of data, since human language has a very complex structure and is highly ambiguous. This research topic has been around for more than 50 years and has enabled computers to handle many complex tasks related to natural language that were believed to be impossible in the last century [1].

Finding relevant articles in today's vast archives of scientific data has become a difficult task. One way to speed up this task is to tag the articles subject in order to ease the search process [2]. In this project we will analyze the title and abstract of a set of research articles and try to predict the topic of the article. For that, we will develop a classification machine learning model capable of such prediction and handling data in natural language. In addition, we will infer the 10 most frequent words in each of the topics of the articles.

2 Problem to Solve

The NLP dataset[1] we have chosen for this project contains the tittle and the abstract of different research articles, as well as the class labels (topic of the article). The problem to be solved is a multilable classification problem in which the research topics can be classified in one of the following fields: Computer Science, Physics, Mathematics, Statistics, Quantitative Biology and Quantitative Finance. Moreover, we have plotted the frequency of the 10 most common words in each of the fields.

3 Experiments done

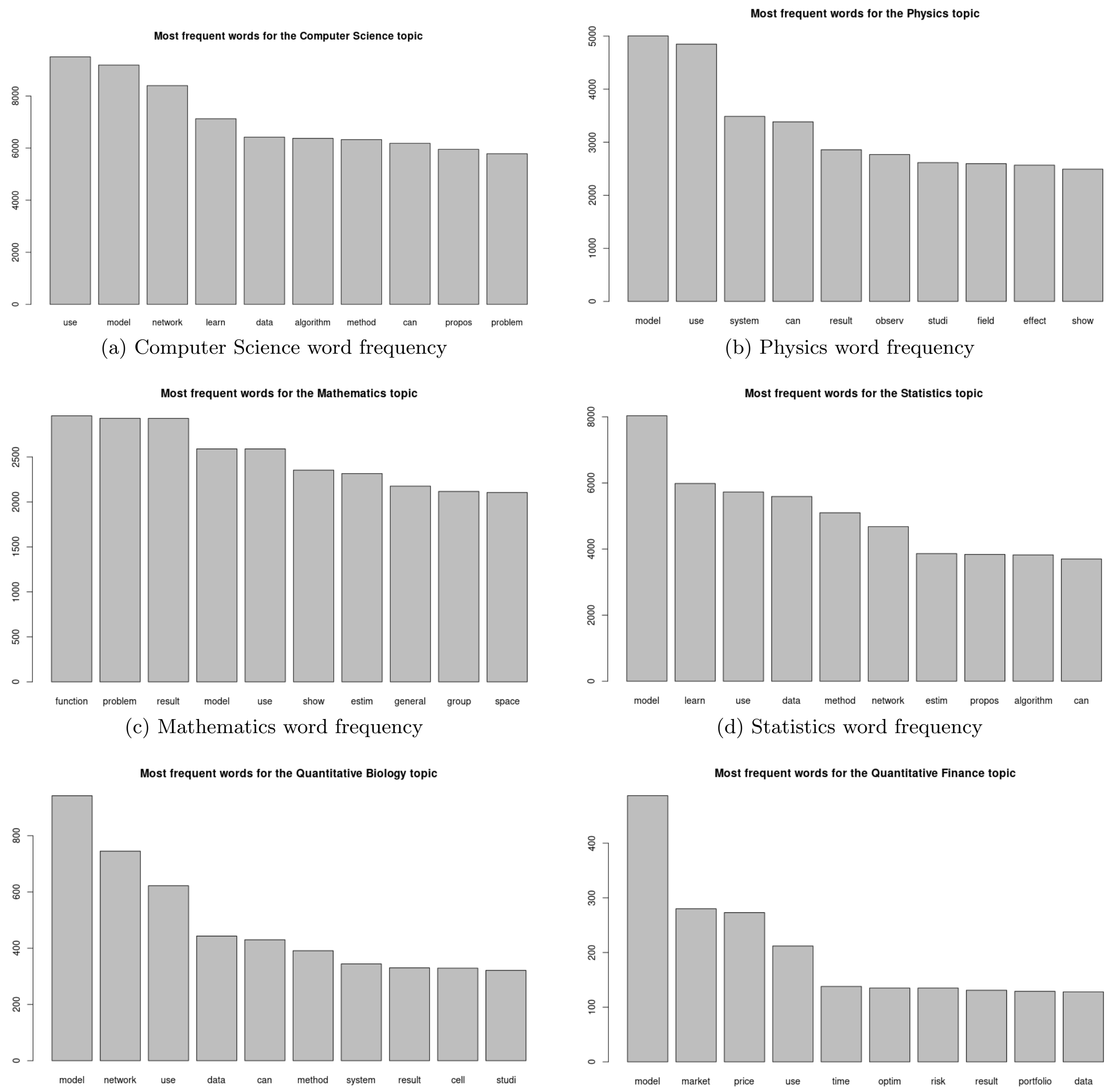
In this section we are going to explain the experiments done with the dataset and also expose and discuss the obtained results.

3.1 Word Frequency

First, we computed and plotted the histogram of the 10 most frequent words in each field, to see what the trend was. These frequent words will play an important role in the ML classification algorithm, as they will give information about what topic the article belongs to.
In Figure 1 it can be seen that words such as cell clearly belong to the biology science field, and market

[1] https://www.kaggle.com/vetrirah/janatahack-independence-day-2020-ml-hackathon

or price to finances. Nevertheless, there are also words that appear with high frequency in all the topics, like model or use, meaning they will not give useful information to our classification system.



(a) Computer Science word frequency

(b) Physics word frequency

(c) Mathematics word frequency

(d) Statistics word frequency

(e) Quantitative Biology word frequency

(f) Quantitative Finance word frequency

Figure 1: 10 most frequent words for each topic

3.2 NLP Classification Problem

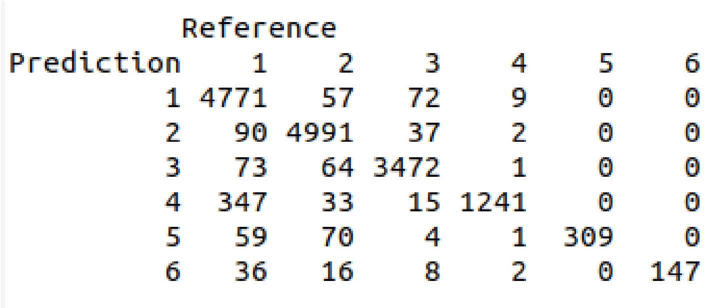
3.2.1 Data Preprocessing

In order to implement a machine learning algorithm, the data had to be preprocessed and cleaned [3]. First, to facilitate the classification problem, we eliminate the instances of the articles that belong to more than one topic and we unite the title and the text columns in the same variable. We also create a

common target variable based on what the actual topic of the article is.
We then proceeded with the data cleaning. We mainly remove unwanted elements and space, convert all text to lowercase, remove punctuation, remove stopwords, and perform stemming.
Finally we create a document term matrix and divide the instances into a training and test set to get the data ready to implement a ML algorithm. The classification model that we have used is the Random Forest algorithm.

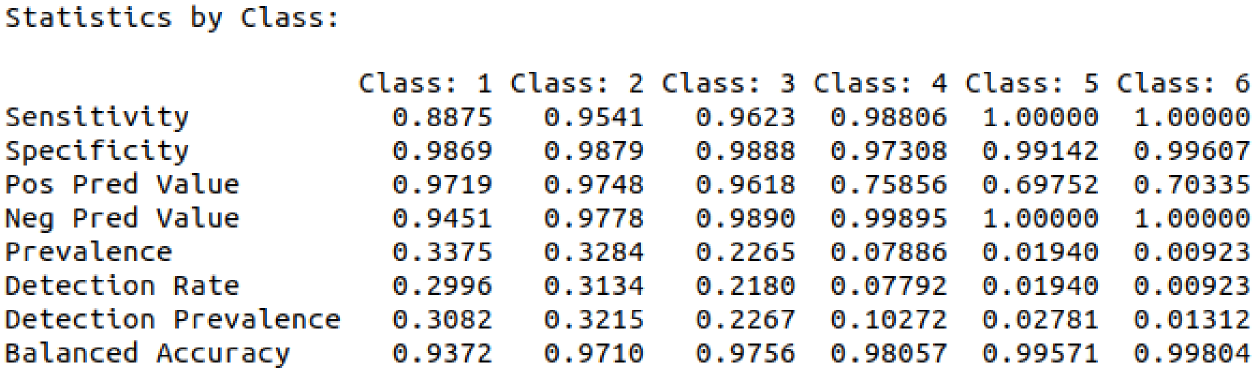
3.2.2 Results and Analysis of the Results

As can be seen from the figures below, the random forest has efficiently classified most of the article topics. The classes 1, 2, 3, 4, 5 and 6 represent respectively Computer Science, Physics, Mathematics, Statistics, Quantitative Biology and Quantitative Finance. We have obtained an accuracy of 0.9375 and a Kappa statistic of 0.9144.



```
          Reference
Prediction    1    2    3    4    5    6
         1 4771   57   72    9    0    0
         2   90 4991   37    2    0    0
         3   73   64 3472    1    0    0
         4  347   33   15 1241    0    0
         5   59   70    4    1  309    0
         6   36   16    8    2    0  147
```

Figure 2: Confusion Matrix



```
Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6
Sensitivity            0.8875   0.9541   0.9623  0.98806  1.00000  1.00000
Specificity            0.9869   0.9879   0.9888  0.97308  0.99142  0.99607
Pos Pred Value         0.9719   0.9748   0.9618  0.75856  0.69752  0.70335
Neg Pred Value         0.9451   0.9778   0.9890  0.99895  1.00000  1.00000
Prevalence             0.3375   0.3284   0.2265  0.07886  0.01940  0.00923
Detection Rate         0.2996   0.3134   0.2180  0.07792  0.01940  0.00923
Detection Prevalence   0.3082   0.3215   0.2267  0.10272  0.02781  0.01312
Balanced Accuracy      0.9372   0.9710   0.9756  0.98057  0.99571  0.99804
```

Figure 3: Statistics by Class

In Figure 2 we can see how the data is unbalanced in terms of the Biology and Finance topics, as there are significantly fewer articles belonging to these topics than the rest. However, we can also see that these are the best classified fields. This is probably due to the fact that they have the most specific words for their field, as we have seen in Section 3.1.

4 Conclusions

Taking everything into consideration we can conclude that we have managed to successfully classify the topic of an article with high performance (accuracy > 90%), based solely on the title and abstract of the work. However, the Random Forest classification algorithm takes a lot of time to compute and in a larger

dataset this could translate in an huge hindrance.
We have also seen that words with high frequency for a specific topic play an important role when classifying the field of an article, but also if they are frequent in many fields they do not help in solving the problem.
The link to the code on Github developed to solve this problem is:
`https://github.com/Julia-upc/NLP-Intelligent-Systems`

5 Further Research

There are several additional changes and improvements that could be made to the project if we had more time and resources. Some examples would be:

- Develop a model capable of classifying articles belonging to more than one topic.
- Oversampling the data to adjust the classes distribution in order to solve the unbalanced data issue of the minority classes.
- Try different classification algorithms to see which one provides the most accurate metrics.
- Develop a more efficient algorithm in terms of computational time.

6 References

[1] Jason Brownlee. What is natural language processing? `https://machinelearningmastery.com/natural-language-processing/`, 2017.

[2] Tamil Nadu Chennai. Nlp on research articles. `https://www.kaggle.com/vetrirah/janatahack-independence-day-2020-ml-hackathon`, 2020.

[3] PLURALSIGHT. Machine learning with text data using r. `https://www.pluralsight.com/guides/machine-learning-text-data-using-r`, 2019.