Text Classification for Serbian Science Journals

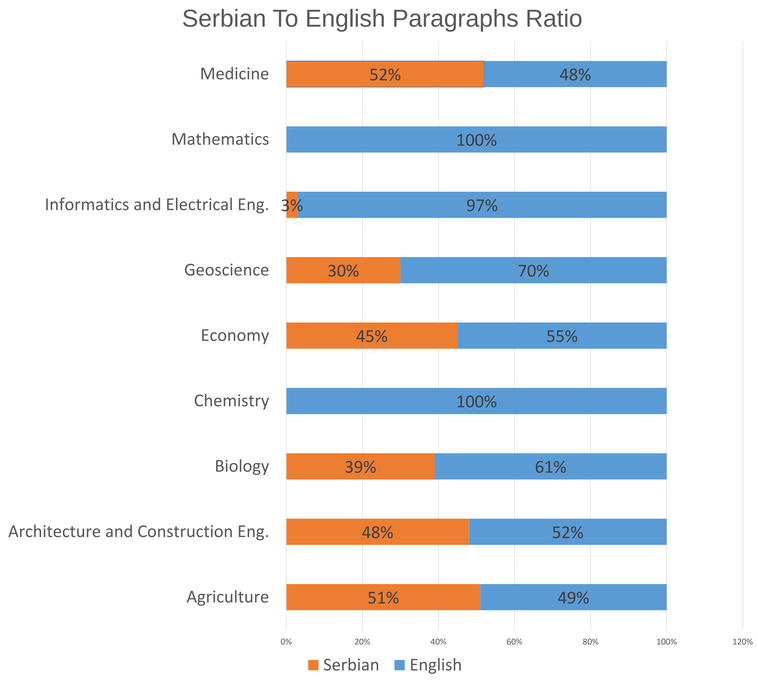
Author: Miloš Đurić

Introduction

The goal of the project is to see how different text classification algorithms classify text paragraphs from Serbian science journals into different categories, depending on their content.

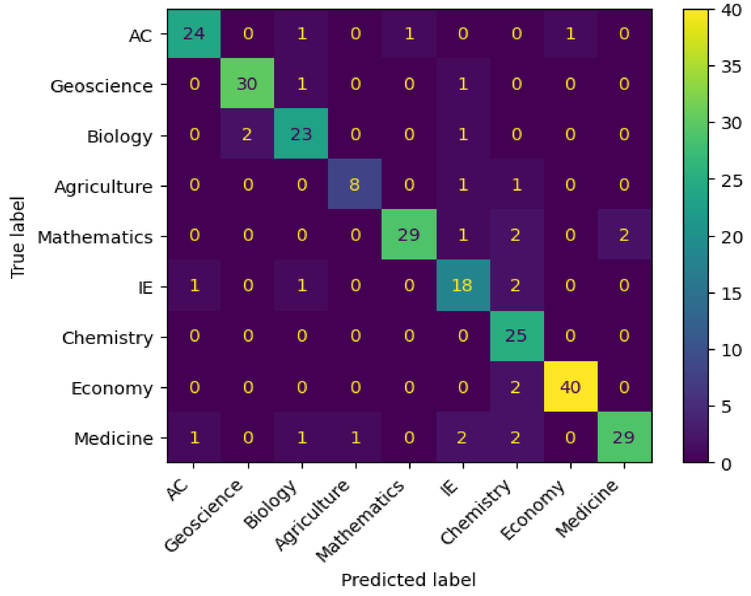
Data

Data used for this project was made using an online archive of scientific journals called “Srpski Citatni indeks”[1]. It contains 1700 paragraphs split into 9 categories. Each paragraph is labeled whether it is in written in Serbian or English. Data is divided 70% in training set, 15% in validation set and 15% in test set



Methodology

- Data is classified using Multinomial Naive Bayes, K-Nearest Neighbors and Support Vector Machine algorithms
- Algorithms are used together with Count and TF-IDF Vectorizers
- Text is filtered by removing insufficient characters and preprocessed with removing stop words, NLTK’s WordNetLemmatizer (English) and SrbAi’s Stemmer (Serbian) before classification
- Metrics used for evaluation are Accuracy and Macro Average metrics
- Confusion Matrix is made for visual representation of the results



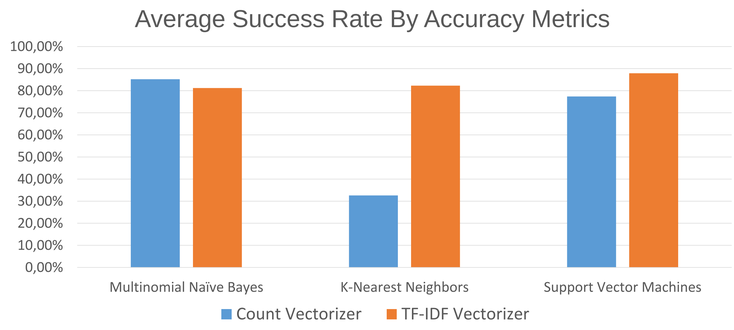
Confusion Matrix Example

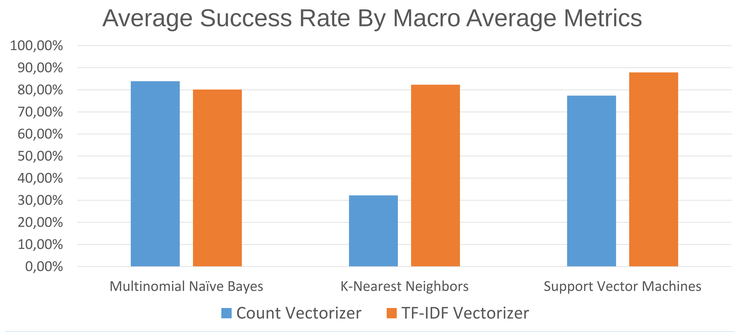
Improving Algorithm Performance

- We use fine-tuning to improve the algorithms results
- Classifiers parameters can be changed, e.g. KNNs weights and SVMs kernel function
- Use of transformers, e.g. PowerTransformer

Results and Discussion

Results show that predominantly SVM with TF-IDF Vectorizer show the best results, while KNN with Count Vectorizer the worst. In general TF-IDF Vectorizer shows greater success rate than Count Vectorizer for all algorithms except for Multinomial NB. Another good performance is shown by KNN with TF-IDF, perhaps because the TF-IDF Vectorizer maps words more adjustably to the KNNs classification than its counter-part. Multinomial NB also shows generally good performance, although its performance doesn’t seem to be greatly affected by the difference in Vectorizers. It seems the use of TF-IDF mapping alongside Support Machine Vectors brings the highest success rate with both accuracy and macro average metrics when it comes to classification of data of this size and vocabulary.





References

- https://scindeks.ceon.rs/