Towards Automatic Classification of EU Projects for Supporting Open Fiscal Data Analysis

Ondřej Zamazal

Department of Information and Knowledge Engineering,
University of Economics, W. Churchill Sq.4, 130 67 Prague 3, Czech Republic
ondrej.zamazal@vse.cz

Abstract. The European union funding for job creation and a sustainable and healthy EU economy and environment is channelled through the five EU structural and investment funds. Although there is EU categorization system for EU projects, EU countries apply their own different categorization systems. Some EU countries already apply European categorization system, but many do not. As a result, many projects in available datasets are not categorized using the European categorization system which hinders straightforward fiscal analyses. The long-term goal of this work is to support an open fiscal data analysis by an automatic classification of EU projects using a machine learning classifier.

Keywords: EU projects, Open Fiscal Data, RDF Code list, Classification

1 Introduction

The European union (EU) funding for job creation and a sustainable and healthy European economy and environment is channelled through the 5 European structural and investment funds (ESIF). Although there is the European categorization system for EU projects, EU countries apply their own different categorization systems. Some EU countries already apply European categorization system, but many do not.

The Open Knowledge Foundation Deutschland (OKFD)[1] gathers data about EU projects published by responsible authorities (often in PDF), clean and share them via the GitHub in different formats such as CSV, XSLX or JSON. During data processing information about EU projects are aligned with one fiscal data model of Open Spending.[2] Recently, OKFD published the *full dataset* described in documents [1] and [2] for period 2000 to 2020 having 2.7 milion of projects.[3] In all, there are 113.446 projects from the latest period 2014-2020 out of which only

[1] https://www.okfn.de/en/

[2] https://github.com/os-data/eu-structural-funds/blob/master/specifications/fiscal.schema.yaml

[3] For this work we used the dataset from the 21st of April 2017. Up-to-date numbers are available at the supplementary web page, http://owl.vse.cz:8080/ISWC2017/

7989 projects (7%) have been categorized into the EU categorization system (intervention code) valid for 2014-2020 period. Although the analysis of the dataset is already enabled by the OS Viewer tool,[4] additional categorization information would enhance straightforward fiscal comparative analyses. The motivation of this work is to support an open fiscal data analysis by EU project classification into the EU categorization system 2014-2020 using a machine learning classifier.

2 The Gathered Data and the Approach

The EU categorization system, having 123 categories, is available in the official EU documents as well as in the tabular form from the *Data for research* web page.[5] We extracted the categorization system for 2014-2020 as the RDF code list[6] where the items contain English labels and are structured in the taxonomy.

The extent to which the European projects in the full data set are described varies a lot: project name, funding amount, funding period, beneficiary name, project description etc. In this work we only focus on a lexical description of the projects; particularly we consider attributes such as *beneficiary name*, *project name* and *project description.* For projects representation we use the bag of words approach where a vector has as many attributes as a number of all unique terms (after removing of stop words) and we count a frequency of a term within the given project. As a result we have data of a high dimensionality. In order to cope with a high dimensionality we chose the machine learning algorithm suitable for such a data [4]. According to our initial testing (see the web page), we decided for the implementation of the classifier based on SVM (Support Vector Machines) with a linear kernel which uses the LibSVM Java library.[7]

In order to gather the training data we first found projects having an intervention code in the dataset. Second, since we needed to unify the semantics of projects' lexical description, we translated all words into the one natural language. We applied the translation into English since translation into this language has usually the best performance and labels in the RDF code list are also in English. We used the *Microsoft Translator API*[8] since this API is available for free up to 2 millions characters per month and there is a sufficient coverage of European languages. As an alternative to natural language translation we experimented with a word sense disambiguation based on Babelfy [3]. Third, in order to obtain only unique projects we deduplicated them based on their lexical description and intervention code. For translated data we arrived at 5269 unique project descriptions where countries are distributed as follows: 3238 Germany, 1430 France, 8 Malta, 198 UK and 395 Greece. For disambiguated data

[4] http://subsidystories.eu

[5] http://ec.europa.eu/regional_policy/sources/docgener/evaluation/data/categorisation_2014_2020_mapping.xls

[6] https://github.com/openbudgets/Code-lists/blob/master/EUcategorization/2014-2020/2014_2020_intervention_fields.trig

[7] https://www.csie.ntu.edu.tw/~cjlin/libsvm

[8] https://www.microsoft.com/en-us/translator/translatorapi.aspx

we arrived at 3631 unique project descriptions where countries are distributed as follows: 1864 Germany, 1187 France, 8 Malta, 195 UK and 376 Greece. Training data are unequally distributed to target classes (intervention codes). Distributions for translated and for disambiguated data are available on the supplementary web page.

3 Preliminary Experiments

In order to evaluate the SVM classifier we have experimented with three different settings with regard to the preparation of training and testing data. Due to the fact that target classes are distributed unequally in the gathered data (and there is no any instance for 32 target classes out of 123), during preprocessing we performed *oversampling*, i.e. some randomly selected instances are duplicated, and *undersampling*, i.e. some randomly selected instances are removed in order to receive 100 training instances per target class.[9]

For training data the experiment A considers only target classes having more than 30 instances, the experiment B considers all target classes having at least one instance, finally the experiment C considers all target classes. For target classes which do not have any training instance we used words in labels from the RDF code list of the categotization system. For testing data the experiments A, B and C include a half of instances from target classes having at least one instance to be used as testing data. Training and testing data are always disjoint.

Table 1. Experiments A, B, C for translated and disambiguated data where Tr means training data, Ts means testing data, # cl. means a number of classes and # in. means a number of instances.

Experiment Translation	**Tr: # cl. (# in.)**	**Ts: # cl. (# in.)**	**Precision**
A	22 (2200)	70 (1691)	.727
B	91 (9100)	70 (1691)	.754
C	123 (12300)	70 (1691)	.762
Experiment Disambiguation	**Tr: # cl. (# in.)**	**Ts: # cl. (# in.)**	**Precision**
A	21 (2100)	71 (1208)	.495
B	91 (9100)	71 (1208)	.517
C	123 (12300)	71 (1208)	.514

Results for the translated and disambiguated data are in Table 1. Regarding translated data from the results we can see that the classifier performance slightly increases from the experiment A to the experiments B and C (from the precision of .727 to the precision of .762).[10] While we could expect that the classifier model which classifies to more target classes (ceteris paribus) would lead to

[9] For training we experimentally used 100 instances, but we want to inspect the influence of this paramater on the classification performance in future.

[10] The precision was averaged from three runs. The demonstration program, EUProjectsClassifier, for each variant is available at the supplementary web page.

lower performance (since it is more complex task) the experiment results show that the performance is better. The classifier in the experiment A provides the classification into 22 most frequent target classes and thus instances of those 22 target classes dominate in the testing data. Although this setting is in favour of the experiment A, the performance in the experiments B and C show that a growing number of target classes is successfully associated with a modest increase of the precision thanks to the fact that the classifier in the experiments B and C can classify additional instances[11] compared with the experiment A. By comparing results for translated and disambiguated data we can see that the performance of classifiers using translated data is better by approximately 20 % than classifiers using disambiguated data.

4 Conclusions and Future Work

This work aims at classification of EU projects in order to support straightforward fiscal analyses. Performed experiments showed that the approach with natural language translation overperformed the approach with disambiguation using Babelfy. Although the results of preliminary experiments are promising, the testing data represent a limited portion of EU projects. Thus, for our near future work we plan to further evaluate our approach using randomly selected and newly annotated testing data. In our ongoing work, we randomly selected EU projects (209.037 projects regardless the funding period) having a non-zero length of the project description and not having the EU categorization. We separated those EU projects into different groups according to the length of their project descriptions and we assigned them to domain experts to annotate them. The initial annotations point out the difficulty of such a task since preliminary inter-annotator agreement is about 25%. Besides this ongoing work on the preparation of the EU projects classification gold standard and the subsequent evaluation of our approach on top of that we further plan to build a web based application for an online classification of given EU project based on a lexical description where the top-n predicted categories will be graphically indicated in the RDF code list visualization.

The work has been supported by the H2020 project no. 645833 (OpenBudgets).

References

1. SubsidyStories.eu. Dataset Descriptions. https://tinyurl.com/ycffsubg. 2017.
2. SubsidyStories.eu. Methodology & Variables. https://tinyurl.com/y7errfx5. 2017.
3. Moro A., Raganato A., Navigli R. Entity Linking meets Word Sense Disambiguation: a Unified Approach. In: Transactions of the Association for Computational Linguistics. 2014.
4. Wang W.,Yang J. Mining High-Dimensional Data. In: Data Mining and Knowledge Discovery Handbook. Springer. 2010.

[11] This can be seen from the precision per target class at the supplementary web page where are also available the results from other experiments with different settings.