Standard Industrial Code Classification by Using Machine Learning

Organisation: Statistics Norway

Author(s): Thivyesh Ahilathasan, Tatsiana Pekarskaya

Date: 07.02.2020

Version: 1

1. Background and why and how this study was initiated

This study was initiated to make the registration of new companies in Central Coordination Register (the administrative business register) more efficient and effective. To complete registration, a company sends a description of business activities to the business register where an executive officer finds a correspondent SN2007 (Standard Industrial Classification - SIC) code based on the business description. The classification is made manually. Taking into consideration that there are around 70 000 companies registering every year, the register process is very time consuming and requires improvements. As a solution could be used machine learning. It can either replace the manual work completely or be used as a support solution while classifying.

This study was therefore initiated to investigate the possibility to use machine learning to classify a SIC code. The first phase of the study which is ended now, was focused on testing a machine learning model up against todays' solution.

The study is a collaboration work of Central Coordinating Register, Statistics Norway, Norwegian Labor and Welfare Administration, Norwegian Tax Administration.

2. Data

2.1 Input Data (short description)

The data that were used to train the ML model were historical data:

1. Descriptions of economic activities, which entities provided under registration with currently defined for them SIC codes;
2. ‘Official’ descriptions of codes and keywords.

Two first are coming from our Business Register, the last - from the SIC system version 2007.

The final dataset contains around 1,5 million descriptions of activities and 821 labels.

We are also looking into possibilities of using business names as features.

2.2 Data Preparation

For the PoC the standard text cleaning was done. Firstly, we removed obviously unreliable associations of activities descriptions and code/s. After we removed stopwords (both standard and extension from text analyses), digits (for some cases), punctuation etc. We turned uppercase to lowercase.

Depending on which model was tested, we converted text to numeric form and left it in text form. Fasttext, which was used for the PoC, did not require conversion. For models, which required numeric interpretation of text, it was used term frequency-inverse document frequency (TF-IDF).

2.3 Feature Selection

Choice of SIC code is naturally coming from business activities, which a company has. The activities. Thus as features were descriptions of business activities. In some cases company names can be informative too, that is why we are also testing the use of business names.

2.4 Output data

The output data is the predicted SIC code and its probability.

3. Machine Learning Solution

3.1 Models tried

The models tested are Logistic Regression, Random Forest, Naive Bayes, SVM, Fasttext and Convolutional Neural Network (CNN).

3.2 Model(s) finally selected and the criterion

To choose between models there were used accuracy, precision, recall and f1 characteristics. Fasttext, SVM and CNN provided around the same result and performed better than other models. However, Fasttext outperformed the others when it comes to a question about time spent for training. That is why it was prefered to the rest.

3.3 Hardware used

We run testing on internal servers (Intel(R) Xeon(R) CPU E5-2699 v3 @ 2.30GHz). The server, we are running on, is using 40 CPU's and 200 GB of RAM.

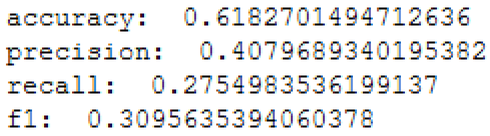
3.4 Runtime to train the model

From 9-30 min for 1 600 000 million training examples and 1 feature, by using Fasttext.

4. Results

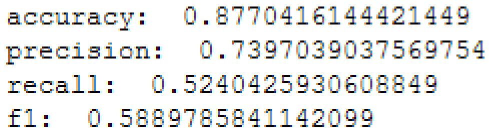
Since we were interested in examining potential of ML classification not only fully replacing manual work but also being a help while classifying, we were looking at partly correct accuracy/f1. Partly correct means that for a unit at least one from five predicted with the highest probability is correct.

The latest results gave:



```
accuracy:  0.6182701494712636
precision:  0.4079689340195382
recall:  0.2754983536199137
f1:  0.3095635394060378
```

And partly correct results:



```
accuracy:  0.8770416144421449
precision:  0.7397039037569754
recall:  0.5240425930608849
f1:  0.5889785841142099
```

Only for 22% of units probability of the first predicted code was higher than 95%.

From Figure 1 we see that the more examples in SIC group are the better model was working.

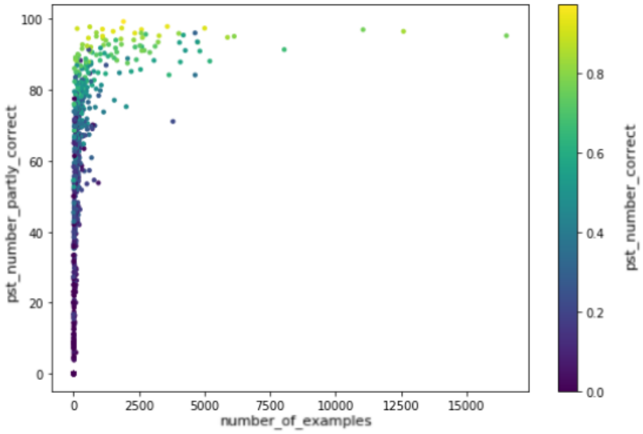


Figure 1. Analysis of results for Fasttext model (Number of examples in a SIC group vs percent of partly correct predicted codes in the groups vs procent of correct predicted codes in the group).

5. Code/programming language

It was used Python language. The python code will be uploaded on GitHub, link coming later.

6. Evolution of this study inside the organisation

In Statistics Norway there are some problems which could be\were solved by using ML. Our study was presented in the company as an example to spread the idea that ML can be used and has a lot of possibilities.

Moreover, people who have been working on the project are the very beginners in ML, who during the project improved their competence in the field.

7. Is it a proof of concept or is it already used in production?

The PoC is already used as a supporting tool by executive officers in Central Coordination Register while registering new entities. For each activity description for executive officers are offered 5 best codes with their probabilities, from which it is faster to make a choice.

7.1 What is now doable which was not doable before?

Registering entities is expected to be faster when using the application that is being developed. The application is expected to result in more consistency and faster registering of entities. The PoC used in production is just for testing.

Over a period of 10 years, the net utility is expected to be between 7 million NOK and 17 million NOK.

7.2 Is there already a roadmap/service journey available how to implement this?

(as many sentences as necessary, as few as possible)

It is under construction.

7.3 Who are the stakeholders?

The stakeholders are Statistics Norway, The Norwegian tax administration, and the Norwegian Labor and Welfare Administration.

7.4 Fall Back

Is a fall back plan in place or planned to mitigate the risk of the ML solution failing in production? Will there be resource left in place to go back to e.g. manual imputation or the use of rule-based scripts? (as many sentences as necessary, as few as possible)

The currently best ML model does not provide us a satisfying level of accuracy to make a full replacement of the manual work. However, it is already used as a supporting tool for manual classification, which saves money. Going back to completely manual imputation is not considered, only improving the model and moving further.

7.5 Robustness

What fail checks are in place or planned to ensure that the ML solution is consistently meeting or exceeding the set gold standard? (as many sentences as necessary, as few as possible)

We are still working on model improvement, since satisfying performance measures were not been reached yet. Thus, it is coming later.

8. Conclusions and lessons learned

ML can be used for text classification but one should have in mind that:

1. Classification for units in some groups might be still low enough to prefer manual correction.

2. It is very important to have reliable units to learn from, both to be able to learn a high performing model and to describe the built model.
3. Even a middle-performing model can provide benefits, saving working time.

9. Potential organisation risk if ML solution not implemented

(as many sentences as necessary, as few as possible)

While not implementing ML will be still only manual work used, which leads both to high time consumption and to misclassified examples.

10. Has there been collaboration with other NSIs, universities, etc?

The study is a collaboration work of Central Coordinating Register, Statistics Norway, Norwegian labor and Welfare Administration, Norwegian Tax Administration.

In HLG-MOS Machine Learning project Statistics Canada, who is working with Fasttext too, has shared their experience with us. Some of their implementations were and still will be tested for our project.

11. Next Steps

We are considering ensembling methods like boosting, bagging and stacking. Adding several models to Fasttext will be relevant then.
The results of the model depend on the dataset the model is being trained on. By eliminating elements that generate noise and correct errors in the dataset, the results can be improved.