Automatic coding of occupation and industry in social statistical surveys: A pilot study

Organisation: Statistics Iceland

Author: Anton Örn Karlsson

Date: October 31st 2020

Version: 2.0

1. Background and why and how this study was initiated

The purpose of the project is to automatize the coding of open answers in social surveys conducted by Statistics Iceland into the respective industry and occupation codes. It is possible to use the data that has been manually coded over the past years by clerical staff working within the institution to train a model that can be applied on future data collected by Statistics Iceland. There are three main reasons for doing this work: 1) Reduce the amount of manual labour needed for coding of occupations and industry in social surveys; 2) Increase the speed of data processing in social surveys with the aim of being able to increase the timeliness of the published data; 3) Provide a proof of concept for the use of machine learning applications in official social statistical production.

The study was initiated and is supported by a grant from Eurostat aimed at increasing the quality of the Icelandic Labour Force Survey (LFS). The study is conducted within the subject matter unit of Labour Force, Living Conditions and Demography.

2. Data

2.1 Input Data (short description)

The input data for the project is data from the Icelandic LFS which has been conducted in Iceland in its current form since 2003. The LFS is a telephone survey where data is collected from a sample selected from the Icelandic population. Open responses on job titles and descriptions of main tasks as well as pre-coded occupational codes are used for training a model for occupational codes. For industry codes the names of enterprises and a description of the enterprises main functions are used as well as a pre-coded industry codes.

The total number of cases from 2003 to 2020 is approximately 160 thousand. In the current state of the project, data from 2017 to 2020 are being used in order to make it more manageable - a total of approximately 19 thousand cases. Also, in the first version of the project, only occupation coding will be examined.

2.2 Data Preparation

In a previous version of the project some text cleaning was attempted. However, because it was both cumbersome (due to the fact that all functions had to be re-written to fit to the Icelandic language) and it did not seem to result in a more accurate model results. Therefore, text cleaning was scrapped in the current version of the project.
The text was tokenized and used as vectors in the final dataset.

2.3 Feature Selection

No.

2.4 Output data

A rectangular dataset where each line is a case in the LFS dataset and each column indicates the ISCO codes available. The values in the set should indicate the predicted probabilities of each case being coded under the ISCO code.

3. Machine Learning Solution

3.1 Models tried

First model to be tested was a naive bayes model and then a simple version of support vector machines. However, it soon became evident that more advanced methods were available which might provide results of higher quality. Therefore, a decision was taken to use deep learning models for training.

3.2 Model(s) finally selected and the criterion

The current final model (which will be reviewed and reworked) was a keras model with three dense layers. The analysis was done by using R with the keras and tidyverse libraries. For the model the rmsprop optimizer was used and the loss function was categorical cross entropy.
The model will be reworked in the coming weeks with the aim of arriving at an accurate final model.

3.3 Hardware used

Intel core i5-7200U, 2.7 GHz.

3.4 Runtime to train the model

A couple of minutes.

4. Results

Figure 1 shows the performance of the model over 20 epochs. An epoch is a single iteration over the training data. In this case (as can be seen in chapter 5) 512 samples are selected from the training dataset as a batch for each epoch, these are the samples used for each iteration of the model. As can be seen from figure 1 the total number of epochs were 20.

In both panels of the figure two datasets are compared, the training data (in orange) which is used to fine tune the model parameters to be better able to predict the occupational codes based on the open text data used for the predictions. The validation data (in blue) is used to see to what extent the prediction holds in another dataset – e.g. to examine to what extent the training data has been overfitted. It can be seen that while the accuracy of the predictions based on the training dataset increases after the 15th epoch, the same cannot be seen for the validation data, indicating that there might be some overfitting in the training data.

The top panel of figure 1 shows the loss of the model. The goal is to minimize the loss in the model and the figure indicates that the loss decreases of the epochs. In this case the loss refers to what extent the model is able to predict the true occupation codes in the dataset – the further away from the true codes the model is, the greater the loss. Therefore, the loss should be as small as possible.

The bottom panel in figure 1 shows the accuracy of the model. It is a measure of how often the predicted occupation code equals the original occupation code.

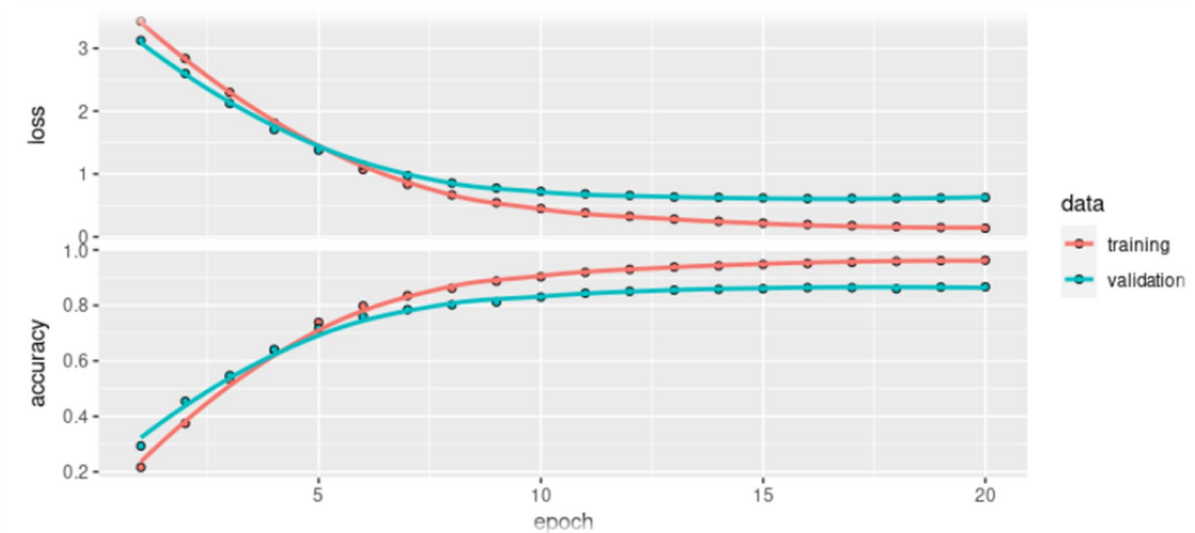

Figure 1. The accuracy and loss of the model over 20 epochs with comparisons between the training and validation dataset.

Evaluation of the model with the test data set:

- Loss 0.6415
- Accuracy 0.8673. Which means that in around 87% of cases the predicted model was correctly able to predict the true two digit occupation code.

5. Code/programming language

The project is being developed in R. The code is still under construction and is heavily influenced by Chollet & Allaire (2018)[1]. Here is a small part of the code, where the model is defined and compiled:

```
train_index <- sample(1:nrow(gg), 0.8 * nrow(gg))
test_index <- setdiff(1:nrow(gg), train_index)

vectorize_sequences <- function(sequences, dimension = 10000) { #dimensions
hérna verða að passa við max_features
  results <- matrix(0, nrow = length(sequences), ncol = dimension)
  for (i in 1:length(sequences))
    results[i, sequences[[i]]] <- 1
  results
}

x_train = vectorize_sequences(text_seqs[train_index])
x_test = vectorize_sequences(text_seqs[test_index])

one_hot_train_labels =
to_categorical(as.numeric(as.factor(gg$kodi[train_index])))
one_hot_test_labels =
to_categorical(as.numeric(as.factor(gg$kodi[test_index])))

model <- keras_model_sequential() %>%
  layer_dense(units = 64, activation = 'relu', input_shape = c(10000)) %>%
  layer_dense(units = 64, activation = 'relu') %>%
  layer_dense(units = 39, activation = 'softmax')

model %>% compile(
  optimizer = "rmsprop",
  loss = "categorical_crossentropy",
  metrics = c("accuracy"))


val_indices = 1:1000
x_val = x_train[val_indices,]
partial_x_train = x_train[-val_indices,]

y_val = one_hot_train_labels[val_indices,]
partial_y_train = one_hot_train_labels[-val_indices,]

history = model %>% fit(
  partial_x_train,
  partial_y_train,
  epochs = 20,
  batch_size = 512,
  validation_data = list(x_val, y_val))

plot(history)
```

[1] Chollet, F. & Allaire (2018). *Deep Learning with R.* Shelter Island, NY: Manning Publications.

```
results <- model %>% evaluate(x_test, one_hot_test_labels)

predictions <- model %>% predict(x_train)
```

6. Evolution of this study inside the organisation

At the moment the study is only being worked on within the department of social statistics – with some involvement from the IT department. Soon it will be presented to a wider audience with the aim that it will inspire further work in the field within Statistics Iceland.

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

The project is both a proof of concept for the use of machine learning within Statistics Iceland and will also be used in production for the LFS. As it is still under development it has not been implemented in the production process of the Icelandic LFS. It is being planned to be implemented in the beginning of 2021. There are no serious challenges foreseen for this implementation – it will simply be an added layer of processing after data has been collected and delivered from the data collection unit.

7.1 What is now doable which was not doable before?

Although the study has not been finalized, it has already established that it is possible to do occupational coding in the LFS much cheaper than before and also far quicker. If the same is possible for industry codes as well, remains to be seen. Therefore, the main added value of using machine learning is increased speed of processing data and decreased cost.

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

The roadmap has not been finalized but in it will entail setting up a virtual machine in the system of Statistics Iceland, then a CI/CD process will be set up for automatic running of the model through the gitlab server of Statistics Iceland. The running of the model will be done by the experts within labour market team in the beginning but over time it will be done automatically.

7.3 Who are the stakeholders?

The main stakeholders within Statistics Iceland are experts working in the labour market team of the social statistics, the data collection unit as well as coders working within Statistics Iceland. The have been informally consulted. A more formal consulting process will soon be conducted as the final model will soon be completed.

7.4 Robustness

In order to ensure the quality of the coded data a portion of the machine coded data will also be hand coded in a blind fashion and then compared in order to estimate the quality of the machine coded data. The final specifications of this has yet to be decided finally.

7.5 Fall Back

Until the machine learning application for automatic coding of occupation and industry have been fully implemented there will still be a possibility to go back to manual coding of the two variables in the LFS.

8. Conclusions and lessons learned

Some of the main conclusion of the project (at the moment) is that a deep learning model can be used for creating machine learning applications for automatic coding. However, the main lessons are the this is not an easy thing to do. A lot of time is need for testing and setting up a model of this type and it is also important to have the relevant IT infrastructure in place. Also, at the moment there is a strong need for an increased awareness of machine learning and machine learning techniques within Statistics Iceland. The current project has advanced the knowledge of machine learning up to a point but there is still an unmet need for more technical competency building and awareness of machine learning and it possibilities within Statistics Iceland. How this will be tackled within the organization remains to be decided.

9. Potential organisation risk if ML solution not implemented

The main risks are that the current labour intensive solution of coding industry and occupation will be continued with the resulting cost for the organization and less than optimal timeliness of the published figures. Also, it might risk that other production processes within official social statistics will not be developed and implemented with machine learning applications where they might be beneficial. Instead, emphasis will be placed on using manual coding.

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

No.

11. Next Steps

The next step of the project is to work further on the model, testing further versions of deep learning modelling techniques and comparing the models in order to find the final version.
Also being planned is to move the model to a virtual machine where more resources can be allocated for the model. In that environment the complexity of the model will be increased by using more cases than currently, test a model for three digits occupational coding and start building a similar model for coding industry as well.
Finally, the model will be moved to Statistics Iceland gitlab server in order to start developing CI/CD applications for the model to be able to run them for automatic coding of data.