Automatic Text Classification of Balance Sheet Data for Uses in Annual National Accounts of Chile

Joaquin Perez-Lapillo

Central Bank of Chile, Statistics Division

jperezl@bcentral.cl

INTRODUCTION

Importance of SUTs and Chilean national accounts

- Supply and use tables (SUTs) play a central role in the System of National Accounts as they provide a coherent and integrated framework for estimating GDP.
- Statistics provided by SUTs enable analysis of the structure of markets and industries as well as studies regarding productivity. One interesting case is the analysis of the structure of the intermediate consumption (IC), i.e. how intensive is the use of goods and services.
- In the national accounts of Chile, output and IC are mainly obtained from business surveys that are expanded to the universe of firms belonging to each industry. The entire process, from data collection to harmonization, takes about 18 months.

New data: IRS balance sheets

- In the last decade, the Chilean IRS has created new strategies to increase the efficiency of tax collection. For the case of balance sheets, the IRS made their digital declaration mandatory for all medium and large-sized companies (25,131 firms) representing about 80% of the Chilean economy annual sales.
- The resulting dataset for the years 2017 and 2018 include 367,547 costs declarations (about 180k rows per year) that can provide a direct contrast to the production functions of each industry. The data is available with a delay of 8 months.
- However, to make the data useful for IC structural analysis, the text descriptor of each cost transaction must be labelled into a standard product classification system.

METHODOLOGY

Compilation of the training set

- The strategy involved the compilation of 10 different datasets already available in the Central Bank of Chile, assuming that the available data can approximate the universe of text to be classified. Some data sources even come from selected balance sheets, meanwhile others origin from transactional data.
- All datasets were manually labelled using the Unique Products Classifier (CUP, in Spanish), which is the official products classifier for Chilean national accounts. The level of detail selected is the 90-class level corresponding to the second level of the CPC 2.1. The compiled training set reached 386,484 total rows.

Text cleaning with NLP methods

- The following steps were taken with the help of natural language processing techniques implemented in the NLTK library for Python: (1) elimination of special characters and digits, (2) tokenization, (3) lower case transformation, (4) stopwords removal, (5) stemming and (6) filtering out tokens containing fewer than 3 characters.

Feature extraction

- The next step involved the transformation of the clean text vector into an efficient numerical matrix. The objective is achieved by the joint application of two methods: (1) word embeddings, using a FastText model that was pre-trained with 1.4 billion words mostly in Spanish that transforms tokens into 300-dimensional vectors, and (2) TF*IDF for weighting relevant words in each row.

Training and evaluation

- To test the learning process, 20% of the observations were isolated. The remaining data is balanced by up/down sampling the target classes depending on their presence. 5 algorithm types were selected and a grid search with 5-fold cross validation was conducted for model selection. Given the class unbalance of the training set, selection is carried out by analysing both accuracy and F1-scores over the unseen data.
- Results show that the best models were RF and SVM, with the first achieving the highest accuracy and SVM the highest F1.

Model	Best hyperparameters	Accuracy (%)	F1-Score (%)
LR	Multiclass	57.5	44.6
RF	300-tree, 50 max depth, entropy	76.2	69.5
NB	Gaussian kernel	41.6	32.5
MLP	2 hidden layers, 200 neurons each, Adam	66.0	56.5
SVM	Polynomial kernel order 3, C=10	73.0	74.9

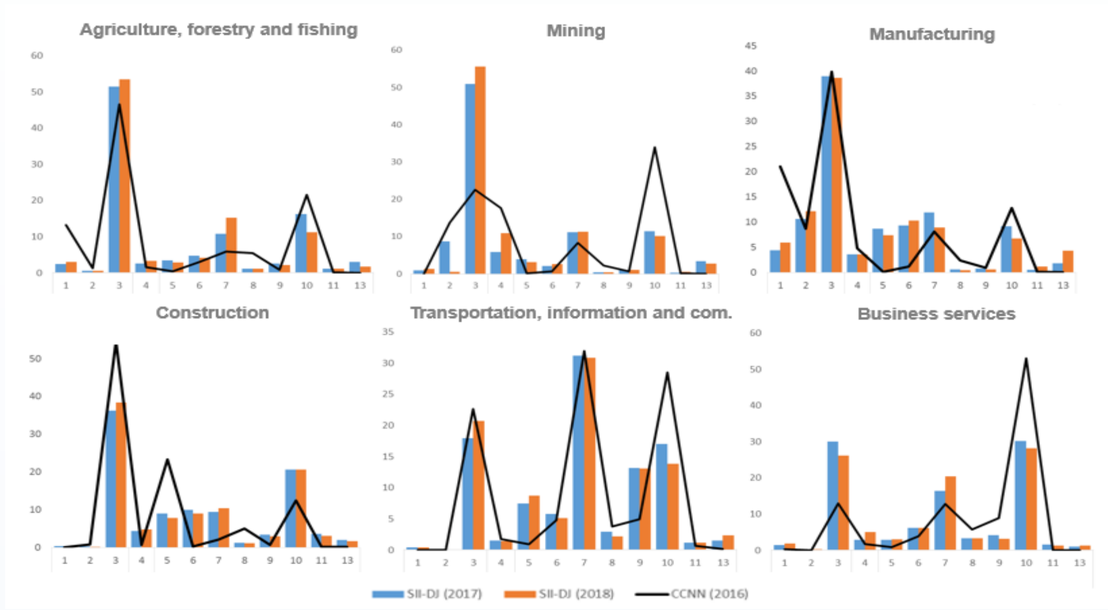
Final evaluation on IRS balance sheets

- The best performing RF and SVM models were finally tested against the unseen cost's dataset provided by the IRS. A small group of experts evaluated a total of 1,000 random observations resulting in RF as the best performing model.

Model	Accuracy (%)
RF	67.2
SVM	66.8

RESULTS

- The figure shows the intensity of use of products as a percentage of the total IC for six relevant Chilean industries. The blue and orange bars were obtained from the predicted product classification provided by the RF model on balance sheet's 2017 and 2018. In addition, official IC structures from the 2016 annual national accounts are shown in black lines.
- In general, results from balance sheet's automatic classification follow the official IC structures. Differences in mining and manufacturing are associated with accounting treatments of work-in-progress and vertical integration of firms that diverge from national accounts concepts.



Horizontal axis ticks: agricultural (1), mining (2), manufacturing (3), utilities (4), construction (5), trade (6), transportation & comm. (7), financial (8), real state (9), business services (10), personal services (11) and not classified (13).

DISCUSION AND CONCLUSIONS

- This work applied text mining techniques and machine learning to exploit a new source of information allowing the extraction of alternative IC structures by the automatic classification of costs into a standard product classifier. These structures are available 10 months before official statistics, representing a great opportunity for the Chilean national accounts.
- Valuable insights emerge when comparing the resulting structures with official data. In cases such as construction, transportation, information and communications and business services balance sheets genuinely suggest a different relevance of inputs that statisticians should consider when elaborating SUTs. In other industries such as mining and manufacturing, the different structures reveal divergences in the treatment of certain accounting items.
- Performance was shown to be acceptable in the context of an industrial application. However, classification errors remain a challenge that could be approached by introducing more manually labelled data into the training set.
- The trained classification model has been useful for other national accounts purposes, such as inventories by product and the validation of firm surveys at a micro-level. Next steps involve the adaptation of the product classification pipeline with other datasets such as the new electronic tax invoices, which impose new challenges due to the volume and variety of the data collected.

REFERENCES

1. United Nations., European Commission., International Monetary Fund., Organisation for Economic Co-operation and Development., & World Bank. (2009). System of national accounts 2008. New York: United Nations.
2. Central Bank of Chile (2016), National accounts of Chile: Compilation of reference 2013. Santiago, Chile.
3. Internal Revenue Service of Chile (2015). Resolution N°112-2015 on "Establishment of form and deadlines to submit the annual affidavit N°1916 and N°1847 with the obligation for medium-sized companies". Santiago, Chile.
4. Aggarwal, C. C. (2018). Machine Learning for Text (1st ed. 2018 ed.). Springer.
5. Jurafsky, D. and Martin, J. (2008). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.
6. Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. (2020). Text Classification Algorithms: A Survey. ArXiv: 1904.08067v5 [cs.LG].
7. Bojanowski, P., Grave, E., Julin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. ArXiv: 1607.04606 [cs.CL].
8. Breiman, L. (2001). Machine Learning, Volume 45, Number 1 - SpringerLink. Machine Learning. 45. 5-32. 10.1023/A:1010933404324.
9. Boser, B., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (COLT '92). Association for Computing Machinery, New York, NY, USA, 144–152. DOI: https://doi.org/10.1145/130385.130401.
10. Murtagh, F. (1991). Multilayer perceptrons for classification and regression, Neurocomputing, Volume 2, Issues 5–6, 1991, Pages 183-197, ISSN 0925-2312.
11. Central Bank of Chile (2018). Supply and use tables, annual national accounts 2016. Santiago, Chile.