Natural Language Processing in Policy Evaluation: Extracting Policy Conditions from IMF Loan Agreements

Joakim Åkerström[1] **Adel Daoud**[2,3,4] **Richard Johansson**[1]

[1]Department of Computer Science and Engineering, University of Gothenburg, Sweden
[2]Centre for Business Research, Cambridge Judge Business School, University of Cambridge, UK
[3]Harvard Center for Population and Development Studies, Harvard University, USA
[4]The Alan Turing Institute, London, UK
gusjoake@student.gu.se, adaoud@hsph.harvard.edu,
richard.johansson@gu.se

Abstract

Social science researchers often use text as the raw data in investigations: for instance, when investigating the effects of IMF policies on the development of countries under IMF programs, researchers typically encode structured descriptions of the programs using a time-consuming manual effort. Making this process automatic may open up new opportunities in scaling up such investigations.

As a first step towards automatizing this coding process, we describe an experiment where we apply a sentence classifier that automatically detects mentions of *policy conditions* in IMF loan agreements written in English and divides them into different types. The results show that the classifier is generally able to detect the policy conditions, although some types are hard to distinguish.

1 Introduction

In the social sciences, evaluating policies often relies on text. What is the effect of a high-ranking politician's tweet on Wall Street? What is the impact of a new economic treaty on trade between nations? What part of the treaty or the tweet induced the relevant effect? These types of policy evaluation questions often require that researchers identify the relevant text passages in large corpora.

Currently, many researchers in these fields devote considerable amounts of resources to hand-coding the relevant passages of the entire corpus of interest (King et al., 2017). For example, social scientists have recently devoted much attention to identifying the impact of macroeconomic policies. These policies affect a population's living conditions both in the short and the long term. The International Monetary Fund (IMF) has since the 1980s been involved in setting the macroeconomic policy space for many countries. IMF's programs contain many different policies, where some might be considered more effective than others. Researchers have therefore sought to compile structured databases identifying what policies each IMF program contains. However, this requires that researches sift all these IMF policies by going through the documents of about 880 programs, between 1980 and 2014, that have been implemented in about 130 countries, and qualitatively hand-coding them (Daoud et al., 2019; Kentikelenis et al., 2016; Vreeland, 2007).

Accordingly, combining qualitative coding to guide a machine-learning powered natural language processing (NLP) tool to operate on large textual data will likely produce large benefits for the social science community. In this paper, we carry out an experiment that investigates the feasibility of developing such a system. We use the IMF research domain as a case study to evaluate the efficacy of our method.

2 Background and Related Work

Textual datasets are often used in investigations in the social sciences, but such investigations typically rely on manual qualitative coding, which is not only labor-intensive but also has the risk of introducing a methodological bias. The principles of grounded-theory has spurred ethnographic and other qualitative research. These principles aim to guide in building social science explanations from the meaning of a corpus (Strauss and Corbin, 1998). Often, this approach does not aim to build systematic coding procedures that are meant to be used in quantitative research. A spin-off of this qualitative methodology, however, called *content analysis*, addresses this gap (Evans and Aceves,

Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 316–320
Turku, Finland, 30 September – 2 October, ©2019 Linköping University Electronic Press

2016). A variety of content analysis has been used to produce databases. Two or more researchers are set to the task of implementing a coding schema interpreting the text and coding it up by hand. Using multiple coders help in estimating inter-coder reliability metrics for qualitative validation. Because it is labor-intensive, content analysis suits smaller-sized corpora.

However, with the rise of larger corpora, the need for automatic content analysis has emerged. This has led to a number of methodological innovations in the overlap between computer science and social science. For example, unsupervised machine learning methods such as topic modeling are often used for various social science problems (Daoud and Kohl, 2016; DiMaggio et al., 2013; Meeks and Weingart, 2012). These unsupervised methods help reduce the dimensionality of the data, but they are unsuitable when there is a clear outcome target – policy text – the researchers desire to code. For these task a number of supervised machine learning methods have been proposed (Grimmer and Stewart, 2013; King et al., 2017). Yet, although a combination of NLP and machine learning are on the rise in computer science, they have yet to fully reach their potential within a social science audience. One way of demonstrating the potential of applying NLP techniques in the social sciences is to evaluate these methods in a real application: extracting policy conditions from IMF reports.

So far, we are aware of no previous work where automated NLP methods have been applied to compile IMF policies from program documents. Most of the research uses qualitative content analysis (Kentikelenis et al., 2016). Recent approaches have been based on a combination of content analysis and a dictionary method to identify IMF food and agricultural policies (Daoud et al., 2019). Some unsupervised methods, mainly different types of topic models, have been applied to the sister organization of the IMF, namely to World Bank, to identify overarching topic changes over time (Moretti and Pestre, 2015).

3 Data and Implementation

3.1 IMF reports

The corpus used in this investigation consists of loan agreements between countries and the IMF, all written in English. These agreements form the policy foundation for the IMF and the recipient government. These agreements outline the macroeconomic problems that the country is facing as well as what the IMF expects from the recipient government. These expectations are defined as a set of policy conditions. The conditions are typically outlined at the end of the loan document.

3.2 Annotation

A team of researchers have coded the policy conditions qualitatively using content analysis principles (Kentikelenis et al., 2016). Two researchers coded all of these policy documents resulting in over 54,000 individual conditions in about 880 programs over the 1978–2016 period. When they assigned conflicting codes, these issues were discussed and resolved by consensus. After all the polices were coded, the next step was to categorize all these individual conditions into overarching policy categories. The categories we consider here are *policy area*, such as finance or environment, and *policy type*, such as benchmarks, performance criteria, etc. This qualitative hand-annotated data provides the input to our supervised training.

3.3 Preprocessing

The IMF documents are stored as PDF documents, some of which required scanning and OCR. The documents go back to the late 1970s, and the quality of the OCR'd text is slightly lower in the earlier documents. Finally, all the documents were converted into plain text using the `pdftotext` tool.

The documents were then scanned to extract the text pieces that matched exactly with the hand-annotated instances. When a text piece consisted of two or more sentences, it was split up to create multiple examples of exactly one sentence each. We did not consider text pieces below the sentence level. Furthermore, all tokens were lowercased and all numeric and punctuation symbols were removed. Stop words were not removed.

3.4 Building the Classifiers

We formalized the extraction of policy conditions as a classification problem on sentences. In the simplest case, the classifier just spots the mentions of policy conditions among a document's sentences. We also extended this basic approach to two different multiclass scenarios, where the policy conditions are subdivided in different ways.

We implemented the classifiers using the scikit-learn library (Pedregosa et al., 2011). The classifiers use a tfidf-weighted feature representation based on n-grams of size one and two, without any feature selection. The classifier is a linear support vector machine with L_2 regularization and a regularization term of 1.0. A one-versus-rest approach was used for multiclass classification. Preliminary experiments using a classifier based on BERT (Devlin et al., 2019) were less successful.

4 Experiments

We carried out a number of experiments to see how well the classifier retrieves policy conditions from the IMF loan agreements, and how well different types of policy areas and policy types can be distinguished.

4.1 Finding Policy Conditions

In the first experiment, we investigated the model's capability of finding mentions of the policy conditions in the documents. This task was framed as a binary classification task where annotated text pieces were treated as a positive set while non-annotated pieces constituted a negative set. The negative examples were subsampled in a random fashion to create a balance of 20% positive examples.

We partitioned the data into training sets and test sets in two different ways. In the first case, we wanted to see how well the classifier generalizes between different countries. We assigned the documents corresponding to 80% of the countries into the training set while the remaining documents were placed in the test set. In the second case, we instead considered the question how well the classifier generalizes to newer documents; in this case, we used the oldest 80% of the documents as the training set.

The classifiers were evaluated using precision–recall curves and average precision scores (AP) to see how well they perform for different classification thresholds. Figure 1 shows the curves and AP scores obtained for the country-based and time-based partition schemes, respectively. It is readily apparent that in both cases the model outperforms a random-guess baseline, which would give an AP score of about 0.20. Furthermore, while the classifier is slightly less accurate when the test set consists of the newer documents, the difference in performance appears to be quite small as indicated by the similar AP scores.

4.2 Classifying the Policy Area

Next, we considered how well the model can classify a text piece as one of several *policy areas*. The data sampling scheme employed in this experiment was similar to the one described in 5.1, with the main difference that the policy area for each example was treated as a target attribute to create a multiclass classification task. Furthermore, the partitioning of the data into training and test sets was performed in a random fashion.

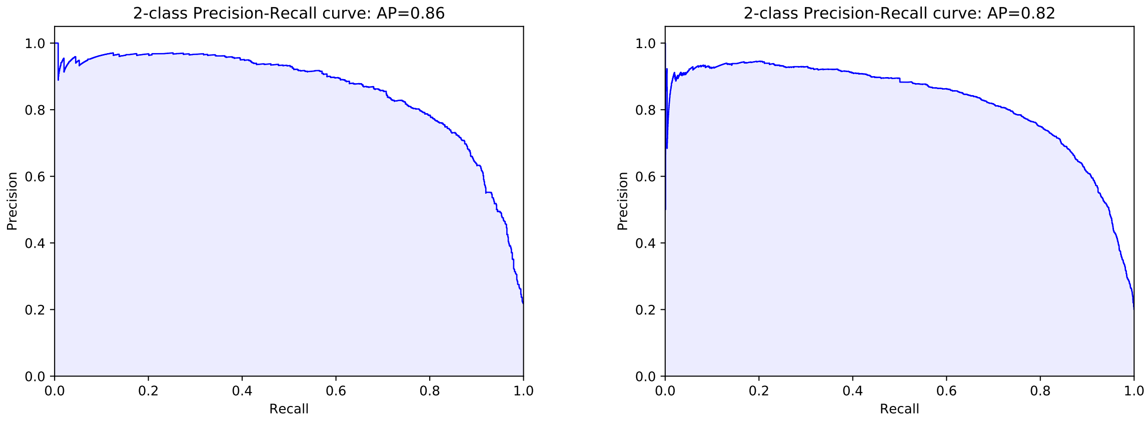
Table 1 shows the precision and recall scores obtained for each individual policy area. The precision scores are consistently higher than the corresponding recall scores. One probable cause for this tendency is the imbalance between positive and negative examples in the training set. Table 1 also shows that the results obtained for some policy areas are remarkably low, with redistributive policies being the most obvious example. The most likely explanation for this phenomenon is the imbalance in the number of training instances per class. Figure 2 compares the F_1 scores for the different policy areas to the number of training examples. While the curve is not perfectly smooth, it is clearly visible that the F_1 score increases quite rapidly with the number of training examples, especially in the lower range of the domain.

Policy area	Precision	Recall
Debt	1.000	0.280
Environment	0.750	0.353
External	0.875	0.491
Finance	0.824	0.618
Fiscal	0.880	0.523
Institutional	0.958	0.354
Labor	0.864	0.520
Redistributive	0.000	0.000
Privatization	0.745	0.522
Revenues	0.863	0.548
SOE	0.918	0.421
Social	0.833	0.469
Other	1.000	0.158

Table 1: Policy area classification scores.

4.3 Classifying the Type of Condition

In the final experiment, we evaluated how well the model distinguishes policy conditions by the *policy type*: indicative benchmark (IB), prior action (PA), quantitative performance criterion (QPC), structural benchmark (SB), or structural perfor-



(a) Training/test split by country.

(b) Training/test split by date.

Figure 1: Precision–recall curves for detecting policy conditions.

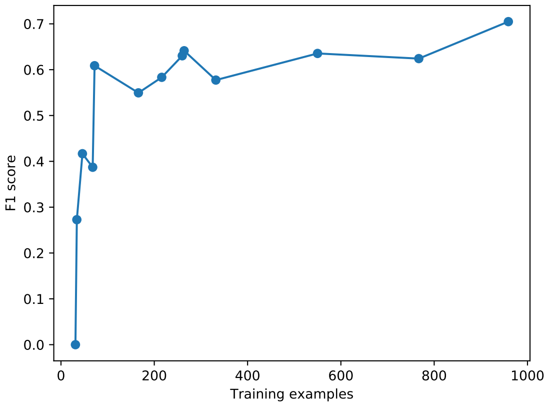


Figure 2: Classification F_1 scores for policy areas as a function of the number of training examples.

Policy type	Precision	Recall
IB	0.955	0.963
PA	0.762	0.535
QPC	0.833	0.269
SB	0.720	0.420
SPC	0.913	0.583

Table 2: Policy type classification scores.

mance criterion (SPC). The partitioning of the data was done in a similar way as in §4.2, with the only difference that policy type was designated as the target attribute. Table 2 shows the precision and recall scores obtained for each individual policy type. As in the experiment on classifying examples according to policy area, the precision scores are consistently higher than the corresponding recall scores and we propose the same explanation for this phenomenon as in §4.2.

5 Conclusions

We have evaluated a sentence classification approach as a supporting technology in a social science research scenario. Our results are promising and show that a straightforward sentence classifier is quite successful in detecting mentions of policy conditions in IMF loan agreements, as well as distinguishing different policy areas and policy types, although the rarer classes are more difficult for our system. This work can be seen as a preparatory effort for the main goal of automatizing coding-based methods in social science, and a more ambitious goal will be to actually apply the classifier in a research scenario and see how the conclusions are affected by the use of an automatic system.

Our use case is just one of many where text processing methods open up new opportunities for changing the way social scientists work with text as research data. Another example is to identify what policies exist around the world: UCLA's WORLD Policy Analysis Center continuously sifts through all the legislation of the world's governments to identify the variation of social, environmental, and economic policies. This includes identifying policies concerning the level of minimum wage, anti-poverty policies, gender inequality, and maternal and child health. These coding procedures require a considerably large team and training to conduct, and this is another scenario where NLP techniques could probably facilitate text-based research in social science.

Acknowledgments

Richard Johansson was supported by the Swedish Research Council under grant 2013–4944.

References

Adel Daoud and Sebastian Kohl. 2016. How much do sociologists write about economic topics? Using big-data to test some conventional views in economic sociology, 1890 to 2014. Technical report, Max Planck Institute for the Study of Societies. Discussion Paper.

Adel Daoud, Bernhard Reinsberg, Alexander E. Kentikelenis, Thomas H. Stubbs, and Lawrence P. King. 2019. The international monetary fund's interventions in food and agriculture: An analysis of loans and conditions. *Food Policy*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. government arts funding. *Poetics*, 41(6):570–606.

James A. Evans and Pedro Aceves. 2016. Machine translation: Mining text for social theory. *Annual Review of Sociology*, 42(1):21–50.

Justin Grimmer and Brandon Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Alexander E. Kentikelenis, Thomas H. Stubbs, and Lawrence P. King. 2016. Imf conditionality and development policy space, 1985-2014. *Review of International Political Economy*, (Online first).

Gary King, Patrick Lam, and Margaret E. Roberts. 2017. Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*.

Elijah Meeks and Scott Weingart. 2012. The digital humanities contribution to topic modeling. *Journal of Digital Humanities*, 2(1).

Franco Moretti and Dominique Pestre. 2015. Bankspeak. *New Left Review*, (92):75–99.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Anselm L. Strauss and Juliet M. Corbin. 1998. *Basics of qualitative research: techniques and procedures for developing grounded theory*, 2nd edition. SAGE, London and New Delhi.

James Raymond Vreeland. 2007. *The International Monetary Fund: politics of conditional lending*. Routledge, Taylor & Francis Group, New York, NY.