Better political text classification using large language models

Information, Redistribution and Financial Regulation conference; Oxford, 1 Oct 2022

Tom Nicholls* Pepper D. Culpepper†

Abstract

Comparative researchers in politics are deeply interested in the ways in which political discourse is conducted for different issues across a wide range of countries, and increasingly use computational methods to classify texts with low cost and high accuracy. Computer scientists are rapidly developing new deep learning models for language tasks, including supervised classification, which are not yet widely used by political scientists. These methods have the potential to improve the accuracy of current bag-of-words methods while also offering the possibility of handing non-English source texts without further work. We present such an improved method for supervised classification using a modern transformer language model, fine-tuned on a large unlabelled corpus and combined with a final softmax layer for probability estimation of category membership. We train the resulting model with hand-labeled data and validate it by analysing a large corpus of news articles on banking. The results show improved classification performance for English-language inputs compared with traditional computational approaches. We also demonstrate the ability to use the same classifier for non-English texts with good levels of classification performance. We suggest that similar methods using large deep learning models are now sufficiently mature for wider adoption by political scientists with primarily substantive, rather than methodological, interests.

Keywords: text-as-data; supervised classification; transformers; deep learning; multilingual analysis

1 Introduction

As political scientists, we frequently want to better understand the language in which political phenomena are discussed. This may be legislative language (e.g. Eggers & Spirling, 2018), news coverage (e.g. Walter & Ophir, 2020), or public discussion on social media platforms (e.g. Majó-Vázquez et al., 2021). Our substantive interest is in the politics of banking regulation: we want to understand the different ways in which banks, bankers, and banking regulations are being discussed in the media before, during, and after

*Department of Communication and Media, University of Liverpool; T.E.Nicholls@liverpool.ac.uk

†Blavatnik School of Government, University of Oxford

the 2008 financial crisis in order to understand the extent to which regulatory changes influence or are influenced by it (see Nicholls & Culpepper, 2021). We are not alone in this; other scholars have explored bank narratives in the contexts of Twitter (Massoc, 2022), news reporting (Lodge & Wegrich, 2011), and parliamentary proceedings (Münnich, 2016). Similar questions arise in the context of migration policy, climate change, military conflict, and every other area of political life. Banking regulation matters, and particularly matters to us, but that is only the tip of the iceberg. Political texts are everywhere, can shed light on a wide range of political scientists' concerns, and remain underexplored.

There are a plethora of techniques for text analysis, but a common and useful approach, and the one we will address in this article, is to classify texts' memberships in theoretically-derived groups. The standard modern toolkit for classifying texts in political science is now computational, relying on relatively small volumes of carefully-coded data being used to train a supervised classification model to classify the bulk of the data. Our common approach is to model each text as a 'bag of words', using only the word frequencies while discarding everything else, and then to train a machine learning model such as a support-Vector machine (SVM) to estimate the probability of each classification for new documents (see Manning et al., 2008). Classifier performance is dependent on the volume of training data given, the complexity of the classification task, and (as when using exclusively human coders) the coherence of the underlying categories.

A new option has recently become feasible with the development of large language models. These models, such as the various kinds of BERT models (Devlin et al., 2019) and other transformers, are complex deep learning systems which also make use of the context of each word in the input text and have been pre-trained on huge quantities of publicly-available text. These can be used for a variety of downstream language tasks: for example, adding an additional softmax output layer then training the model on labelled data produces a supervised classification model. Large language models are potentially capable of greater classification accuracy than systems such as Naïve Bayes and SVM, given that they can both transfer learning from their pre-training data and also make use of the structure of incoming texts. They are, however, not yet widely used by political scientists.

Effective supervised classification becomes even harder when a project requires classifying content in multiple languages, as our standard computational toolkit is designed to work on texts in a common language. A traditional, if unsatisfactory, option is to construct parallel models in each language, with the analyst reading the results across. There is now also an extensive body of work on the translation of texts to a common language (normally English) followed by monolingual analysis (e.g. Courtney et al., 2020; de Vries et al., 2018; Lucas et al., 2015; Reber, 2019), despite reasonable objections from comparativists that it is better to work in the original languages than in translation. A significant practical difficulty with this approach for very large datasets, however, is that the well-studied approach relies on machine translation via a state-of-the-art cloud provider, usually Google Translate or DeepL. This is not a viable strategy for the many researchers with more data than funding, given a cost of around $/€20 per million characters for API access on these services[1]. Publicly available machine translation models also exist (e.g. Fan et al., 2021; Tiedemann & Thottingal, 2020) and can be run locally. It is not clear whether their (less accurate) results are reliable enough to be used for substantive research.

The multilingual problem, too, can potentially be addressed by newer approaches to classification. Some large language models have been trained on texts in many languages. This has important applications in machine translation, but also allows a single model to be trained to do supervised classification in multiple languages in the same way as a single language model: by fine-tuning and training the final classification head of the model using multilingual data,

In this paper, we present and validate a model for supervised classification based on a multilingual transformer model. We are engaged in a large multinational research project on the politics of banking regulation for which the supervised identification of media frames in news coverage across languages is important; our data and classifications are drawn from this motivating example, and feature substantively interesting categories with complex decision-boundaries and relatively low volumes of trained data. We comparatively test two approaches: a multilingual deep learning model with a classifier head, and local machine translation followed by a conventional supervised classifier. We outline the collection and preprocessing of our dataset, the methodological choices made for our two modelling approaches, and comparatively analyse the results of the two strategies for both the traditional (English-only) and multilingual cases. Finally, we offer observations and comments for the wider use of these methods in the future.

2 Data

We use a large dataset (N=1,084,457) of news articles on banks and banking from six countries (Australia, France, Germany, Switzerland, UK, and US) published between 2007 and 2018. These data are substantively interesting for us, as researchers interested in the political economy of banking regulation, but also have a number of useful features for comparative analysis of multilingual methods.

Firstly, the issues raised by international finance are complex, as is the way coverage of the topic is framed in different newspapers. Consequently, the classification problems are decidely non-trivial for human coders, let alone for machines. Secondly, the problem is genuinely international, as the banking system is globally connected and many of the regulatory challenges are handled with international cooperation. Consequently, we can have reasonable confidence that it *is* sensible to attempt classification of content

[1]There are various workarounds to automate translation without an account by abusing the web interface; these are in breach of the terms of service and also somewhat fragile.

across countries and languages using a single model.

We want to assess the widest possible breadth of political discourse on banking. Although we cannot and should not assume that we have an exhaustive collection of everything, the so-called N = all (Jungherr & Theocharis, 2017), we certainly want to move beyond the traditional approach of studying a sample of articles from a handful of elite newspapers in the US (or possibly the US and one or two other rich English-language countries). We want to be able to make the different discourses and framings for any issue, in any country for which data is available, legible to political scientists. Consequently, we have collected data very widely from each of the newspapers we think important in each target country. As the structure of each of the media markets are different, the criteria were different for each, but the sample generally contains the largest and most important national newspapers (including tabloid, broadsheet, and specialist business dailies where possible). We see this as an advantage of this dataset, allowing us to grasp the ways regulatory or political discourse in the press look across countries, which should be a central goal of comparative politics, but has often been too hard.

Our categories for classification are informed by the literature on emphasis framing (Chong & Druckman, 2007; Slothuus & de Vreese, 2010). They were selected for substantive relevance, and are typical of the kinds of complex classifications that are often necessary in practical political communication work. For each of the categories, the decision boundaries are complex, with nuances and exclusions, and are thus a challenge for automated classification. Classification is multi-membership, in that articles can contain zero of these frames, one, or multiple; consequently, we fit separate models for each category rather than a single multi-category classifier. A simplified summary of the categories in the table below[2].

Table 1: Banking article classifications

Classification	Definition
Scandal	Named banks (allegedly) involved in, or allegedly involved in, *misconduct* that involved settlement/charge/fine/claim etc., or a sense that banks are *ripping off* customers.
Regulation	The making of rules for banks or the regulatory or political oversight of banks
Business as usual	The business of banking as a normal economic activity, including earnings, business strategy, the development of new products, and discussions of senior leaders in banks
Financial crisis/bailouts	The events of the financial crisis of 2008 and the bailouts that resulted from them (including the lingering effects years afterward)

[2]The full coding document is in the appendices.

Classification	Definition
Executive pay	The level of executive pay and bonuses, including shareholder votes on (or revolts about) remuneration of senior executives

3 Method

Data were processed and extensively cleaned in R and Python, using the `tm` (Feinerer et al., 2008) and `quanteda` (Benoit et al., 2018) frameworks. As LexisNexis and the other databases can contain multiple copies of each article (normally representing multiple regional editions or different recensions) we de-duplicated articles within each newspaper, using a cosine similarity of 0.65 over a standard bag-of-words vector model as the upper bound for treating articles as unique (see Boumans et al., 2018).

Articles are of wildly varying length, and some cover a range of different topics. Consequently, our unit of analysis is the first one or more paragraphs including the headline and standfirst rather than the whole of the article. We use a modified version of the algorithm used by Gilardi et al. (2020), combining short paragraphs where necessary and aiming for a document length as little as possible above 150 words.

Each of the French and German texts were machine translated to English using the fr-en and de-en pairs from Opus-MT (Tiedemann & Thottingal, 2020), via the EasyNMT library (Ubiquitous Knowledge Processing Lab, 2021).[3]

Because a keyword search on banking related terms includes a large number of articles which are non-bank related (or only feature banks and banking very slightly) we took a two-stage approach (after D'Orazio et al., 2014). For the first stage, a preliminary filter was carried out using a variation of the approach in Benoit et al. (2016): crowd-sourced labelling was used to identify relevant and non-relevant articles, and a linear SVM using tf-idf term features was used to identify and exclude non-relevant articles.[4] The resulting data corpus (N=556,871) is a much cleaner representation of our subject of interest than the original collection, though there are still a number of non-banking articles which have slipped through to complicate our downstream analysis.

For training, 850 random English documents were selected, together with 400 French and 400 German. For each category, articles were double-coded, with disagreements resolved by discussion. The training data was split into separate training/test (75%) and holdout (25%) samples.

[3] This was carried out using a cluster of machines in the University of Oxford's ARC high-performance computing centre, using GPUs. It is beyond the sensible capacity of a single researcher's desktop CPU, but accessible to those with access to modest institutional HPC systems.

[4] Performance was validated using a separate sample of documents coded by the authors, with the French and German content processed in Opus-MT translation. In all cases the performance of both the crowd and the classifier were good but not stellar, with an F1 performance of around 0.8; a full analysis of the merits of crowd-sourced pre-filtering is outside the scope of this paper.

As the category distributions in the data are very unbalanced, top-up samples of likely-positive articles were created for each category. After coding the random sample, a keyword analysis was carried out to identify features associated with positive membership in each category, and a set of English documents were selected for top-up coding using the same procedure as above (scandal: 113 documents; regulation, crisis/bailouts, and executive pay: 250 documents each; business as usual: no top-up). As this sample was non-random, these labels were used exclusively for training/test data, and not for final validation.

For the conventional approach, a number of standard modelling approaches (including non-linear SVM, logistic regression, and random forests) were trialled, with the best performer being a standard linear SVM model. Two sets of linear SVM classifiers were fit using `sklearn` (Pedregosa et al., 2011). In each case the features used for the models were a conventional tf-idf weighted bag of words, with the hyperparameters chosen using an initial grid search over the data for the "scandal" frame. As the models are monolingual, the Opus-MT machine translated texts are used for French and German documents.

The first set of models reflects the common monolingual analysis, by training and validating exclusively using the English-language data. The second set were trained on the (Opus-MT translated) French and German training data too, and validated using each of the three languages' holdout data. This reflects the multilingual case.

Transformer models are trained on extremely large corpora of text from sources such as the web. Nevertheless, they are known to perform better for downstream tasks if the pre-trained weights are fine-tuned, in an unsupervised manner, on texts which are specific to the domain and task (Devlin et al., 2019).

The feature selection and classification parts of a traditional model are somewhat combined here: transformer models both generate an internal vector representation of the text and classify it (using the additional model layer). The complete model is then fitted using backpropogation using the training data.

There are a number of pre-trained BERT examples, none of which are perfect for our application: they are trained on general English corpora rather than on news in general or banking news in particular. We have dealt with this in the standard way using *transfer learning*, taking a standard pre-trained model and fine tuning it for our own problem. This is more effective than using the model as-is, and *vastly* cheaper and easier than attempting to train one from scratch (which requires hundreds of gigabytes of training data and millions of dollars worth of computing time).

In our case, we fine-tuned the XLM-RoBERTA (Conneau et al., 2020) cross-lingual model on all of the documents in our corpus, in their original languages (691,495 English; 178,603 French; and 223,979 German), using the Hugging Face processing framework in Python. Finally, we added a final softmax classification layer, and copied the resulting fine-tuned base classification model to allow it to be trained separately with labeled data for each classification task.

4 Results

Table 2: Classification performance, by language and classification task (macro F1)

Language	Model	Scandal	Regulation	BAU	Crisis/Bailouts	Exec Pay
English (only)	SVM	0.82	0.85	0.72	0.74	0.72
English (only)	RoBERTa	**0.89**	**0.89**	**0.81**	**0.80**	0.72
English	SVM	0.87	0.85	0.78	0.79	0.62
English	RoBERTa	**0.88**	**0.86**	**0.85**	**0.84**	**0.72**
French	SVM	0.79	0.77	0.79	0.75	0.72
French	RoBERTa	**0.85**	**0.79**	**0.85**	**0.81**	**0.77**
German	SVM	0.75	0.75	0.71	0.64	0.72
German	RoBERTa	**0.77**	**0.84**	**0.81**	**0.82**	**0.77**

The XLM-RoBERTa classifier has performed at least equivalently and generally substantially better than the SVM across all classification tasks and all languages. Macro F1 performance gains are between 0 and 9 percentage points in the English-only model (average 4.3) and between 1 and 18 (average 5.1) using the multilingual model.

Improvement has been greatest in the tasks with the most complex decision boundaries[5] and lowest SVM performance. Business As Usual, in particular, is a diverse category covering a wide variety of underlying topics. The Crisis/Bailouts category codes for the aftermath of the 2008 financial crisis and is very context dependent so difficult to code from a bag of words – ideal territory for a more complex language model such as XLM-RoBERTa.

For the SVM models, performance is best on the frames with the clearest lexical markers (scandal and regulation), most challenging on executive pay (which is wildly unbalanced: only around 3% of articles in the sample are positive for this category) and most mediocre for business as usual (which co-occurs with many of the other frames and has the widest definition and the loosest boundaries).

The results also reflect the variations between the different classifications that have been chosen. The usual problems of disambiguating messy categories with complex decision boundaries do not disappear just because of the selection of a more complex (and generally more effective) language model. XLM-RoBERTa

[5]see the coding instructions in the appendix

has access to much greater context about word use than the SVM's simple bag of words, and can apply transfer learning from a truly vast original training corpus. Nevertheless, where there is inherent ambiguity between classes it is not magic.

5 Conclusions

The results show the advantages of large language models for the kind of classification tasks common in political science. Given the same training data, the new XLM-RoBERTa approach outperformed the conventional SVM across the board, providing greater accuracy in classification even for the classic monolingual case. Performance was particularly improved for the trickier classification tasks with fuzzier decision boundaries which were harder to classify for both humans and the SVM.

When working with the multilingual data, the XLM-RoBERTa model had the dual advantages of offering classification performance beyond the SVM baseline and *also* avoiding the expensive and/or time-consuming machine translation of the non-English sources. Offering classification performance beyond the SVM baseline on the English dataset alone, its ability to operate multilingually comes 'for free'. As a result of the transfer learning possible between languages, results nearly as good as English were possible with very low quantities of other-language training data.

The requirements for researchers for running these models are somewhat higher than those for conventional methods, but well within the reach of researchers using computational methods. It is possible, for example, to use deep learning models as a drop-in replacement for parts of an existing classification pipeline using scikit-learn in python, with model performance validated in an identical way. For fine-tuning and training transformer models a suitable GPU is much faster than running on a standard laptop, though XLM-RoBERTa handles a wide range of languages beyond English, French, and German, and other language models pre-trained on particular language combinations are also available.

Work on transformers and other deep learning models is still progressing at a rapid pace and both Opus-MT and XLM-RoBERTa have already been joined by newer and more accurate language models. Nevertheless, the ecosystem is now robust enough that models are fairly straightforward to obtain, train, and use for inference, given suitable hardware. We warmly endorse their wider use by political scientists.

Acknowledgements

We would like to thank Robin Hsieh and Clemence Hautefort for their research assistance, and to all those other colleagues who have given useful comments and advice during this research. We would also like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work. http://dx.doi.org/10.5281/zenodo.22558

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 787887). This paper reflects only the authors' views; the European Research Council Executive Agency is not responsible for any use that may be made of the information it contains.

Appendices

Article selection

Most articles were sourced from LexisNexis. Where LexisNexis data were unavailable, other sources were used including the ProQuest and Factiva databases and by crawling the online archives of newspapers using the RISJbot (Nicholls, 2018). In all cases documents were been selected by keyword. In English, articles were selected containing at least two words with the stem `bank!`. In French a similar search was carried out, `bank! OR banque OR bancaire OR banquier OR banquière OR financier OR financière`. In German, the extensive use of compound nouns required a more complex search strategy, with search terms developed based on analysis of the Bild and Zwanzig Minuten web crawl data: `bank! OR bundesbank! OR citibank! OR commerzbank! OR dekabank! OR deutschebank! OR direktbank! OR geschäftsbank! OR genossenschaftsbank! OR gierbank! OR hypovereinsbank OR immobilienbank OR investitionsbank OR investmentbank! OR krisenbank! OR landesbank! OR mittelstandsbank! OR nationalbank! OR nordbank! OR onlinebank! OR pfandbriefbank OR postbank OR privatbank! OR raiffeisenbank! OR targobank OR volksbank! OR zentralbank! OR nord lb OR bayernlb OR consorsbank OR comdirect OR kantonalbank!`.

Codebook

These categories were derived inductively by looking at unsupervised categorizations of front-page stories, aggregating the stories about banks into politically meaningful categories. Politically meaningful means we excluded general discussions of the direction of the economy or of the stock market, in which banks were included as examples. In each of the categories we selected for analysis, the banks were protagonists, not examples or illustrative cases.

`1. Scandal`

All the articles about scandals that are clearly in this category feature one or more named banks that were involved in, or allegedly involved in, misconduct that involved settlement/charge/fine/claim etc., or a sense that banks are ripping off customers. The lowest of these is a customer complaint, which I have put in the scandal category only when they come in large bulk when they are reported (write-in advice columns where customers note an individual complaint and ask if their bank is allowed to charge them for product x or fee y are not included). Fixing of exchange rates, such as Libor and Euribor, are clearly in. Hidden fees (ATM or credit card or mortgage exit fees), if not found to be illegal, are not a 'scandal'; if a regulator forces fees to be reimbursed by the banks, that does count as scandal/misconduct. If a bank is completely exonerated of charges of some sort in an article, this is also not a scandal. Line of business lawsuits between large financial institutions (e.g. over the interpretation of terms in a bond deed) are not inherently scandalous.

If there is an investigation of a bank for enabling tax dodging by individuals, or if there is an investigation of a bank or its employees for breaking banking secrecy rules, that is a scandal. If individuals are linked to/investigated for tax dodging in another country, and their activity is also tied to a specific bank, that is also in. If there are criminal investigations related to tax evasion and banks, that is a scandal.

`2. Regulation`

These are articles about the making of rules for banks or the regulatory or political oversight of banks. Articles about financial reform, rules about financial stability, liquidity requirements such as Basel III or CRDIV, European banking union, stress tests, or prudential regulators (whether micro-prudential or macro-prudential) fit this category. Articles about taxes specific to banks also fit this category. Articles about backlash against potential or actual regulation fit this category, as do congressional or parliamentary hearings about banks. Articles about banks organizing for political action – to shape regulation – also fit this category. Articles mentioning banking regulators also fit this category. Political oversight includes by ministers, the political opposition and any organised campaign louder than one lone backbencher. Discussions about changing the legal rules of confidential customer information – either to or away from more secrecy – fall into this category.

`3. Strategy and products, executives, and profit/loss - business as usual`

These are articles that deal with the business of banking as a normal economic activity, including earnings, profits and loss, and share price of the bank. Articles about business strategy, entering or leaving particular markets, the development of new products (savings accounts, credit cards), and discussions of senior leaders in banks (including obituaries of senior bankers), as well as their boards, all fit this category. Articles about fees, which are not about alleged misconduct, fall into this category as products. Articles about fintechs (financial services through technology firms), where they discuss these firms competing with banks in certain markets, fall in this category (articles about fintechs without reference to banks fall in none of these categories).

`4. Financial crisis and bailouts`

These are articles that deal with the events of the financial crisis of 2008 and the bailouts that resulted from them. Articles about particular institutions involved – such as Northern Rock or Lehmann Brothers

– are in this category, as are articles that discuss the details of taxpayer bailouts, specific bailout policies such as TARP, toxic assets or the sources of the crisis – too-big-to-fail banks, high risk financial products such as CDOs based on the sub-prime market, or bank borrowing standards that are low because banks don't hold the risks of mortgages – all of these are in. The category is inclusive; if the crisis or bailouts are mentioned in the article more than minimally then it's in, even if it's about lingering effects rather than the ongoing event. Articles dealing only with bailouts from the Eurozone crisis of 2009-10, without making any reference to the global financial crisis, are excluded from this category.

`5. Executive pay`

These are articles about the level of executive pay and bonuses, including shareholder votes on (or revolts about) remuneration of senior executives.

Validation set details

Table 3: Validation set category membership

Language	N	Scandal	Regulation	Business As usual	Crisis/Bailouts	Exec Pay
English	210	31	61	83	30	6
French	200	31	33	89	47	7
German	200	26	58	91	48	7

References

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data. *American Political Science Review*, *110*(2), 278–295. https://doi.org/10.1017/S0003055416000058

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Boumans, J., Trilling, D., Vliegenthart, R., & Boomgaarden, H. (2018). The agency makes the (online) news world go round: The impact of news agency content on print and online news. *International Journal of Communication*, *12*, 1768–1789. Retrieved 11/05/2018, from http://ijoc.org/index.php/ijoc/article/view/7109

Chong, D., & Druckman, J. N. (2007). Framing theory. *Annu. Rev. Polit. Sci.*, *10*, 103–126. https://doi.org/10.1146/annurev.polisci.10.072805.103054

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). *Unsupervised Cross-lingual Representation Learning at Scale*. Retrieved 05/05/2022, from http://arxiv.org/abs/1911.02116

Courtney, M., Breen, M., McMenamin, I., & McNulty, G. (2020). Automatic translation, context, and supervised learning in comparative politics. *Journal of Information Technology & Politics*, *0*(0), 1–10. https://doi.org/10.1080/19331681.2020.1731245

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805

de Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications. *Political Analysis*, *26*(4), 417–430. https://doi.org/10.1017/pan.2018.26

D'Orazio, V., Landis, S. T., Palmer, G., & Schrodt, P. (2014). Separating the wheat from the chaff: Applications of automated document classification using Support Vector Machines. *Political Analysis*, *22*(2), 224–242. https://doi.org/10.1093/pan/mpt030

Eggers, A. C., & Spirling, A. (2018). The shadow cabinet in Westminster systems: Modeling opposition agenda setting in the House of Commons, 1832–1915. *British Journal of Political Science*, *48*(2), 343–367. https://doi.org/10.1017/S0007123416000016

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., & Chaudhary, V. (2021). Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, *22*(107), 1–48.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54. http://www.jstatsoft.org/v25/i05/

Gilardi, F., Shipan, C. R., & Wüest, B. (2020). Policy Diffusion: The Issue-Definition Stage. *American Journal of Political Science*. https://doi.org/10.1111/ajps.12521

Jungherr, A., & Theocharis, Y. (2017). The empiricist's challenge: Asking meaningful questions in political science in the age of big data. *Journal of Information Technology & Politics*, *14*(2), 97–109. https://doi.org/10.1080/19331681.2017.1312187

Lodge, M., & Wegrich, K. (2011). Arguing about Financial Regulation: Comparing National Discourses on the Global Financial Crisis. *PS: Political Science & Politics*, *44*(4), 726–730. https://doi.org/10.1017/S1049096511001351

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, *23*(2), 254–277. https://doi.org/10.1093/pan/mpu019

Majó-Vázquez, S., Congosto, M., Nicholls, T., & Nielsen, R. K. (2021). The role of suspended accounts in political discussion on social media: Analysis of the 2017 French, UK and German elections. *Social Media + Society*, *7*(3). https://doi.org/10.1177/20563051211027202

Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Massoc, E. (2022). *Fifty Shades of Hatred and Discontent - Varieties of Anti-finance Discourses on the European Twitter (France, Germany, Italy, Spain and the UK)* (No. 338). Liebnitz Institute for Financial Research SAFE. Frankfurt, Germany. Retrieved 26/09/2022, from https://papers.ssrn.com/abstract=4008884

Münnich, S. (2016). Readjusting imagined markets: Morality and institutional resilience in the German and British bank bailout of 2008. *Socio-Economic Review*, *14*(2), 283–307. https://doi.org/10.1093/ser/mwv014

Nicholls, T. (2018). *RISJbot: A scrapy project to extract the text and metadata of articles from news websites* (Version v1.1.0). https://doi.org/10.5281/zenodo.1341873

Nicholls, T., & Culpepper, P. D. (2021). Computational Identification of Media Frames: Strengths, Weaknesses, and Opportunities. *Political Communication*, *38*(1-2), 159–181. https://doi.org/10.1080/10584609.2020.1812777

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Reber, U. (2019). Overcoming Language Barriers: Assessing the Potential of Machine Translation and Topic Modeling for the Comparative Analysis of Multilingual Text Corpora. *Communication Methods and Measures*, *13*(2), 102–125. https://doi.org/10.1080/19312458.2018.1555798

Slothuus, R., & de Vreese, C. H. (2010). Political Parties, Motivated Reasoning, and Issue Framing Effects. *The Journal of Politics*, *72*(3), 630–645. https://doi.org/10.1017/S002238161000006X

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT–Building open translation services for the World. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 479–480.

Ubiquitous Knowledge Processing Lab. (2021). *UKPLab/EasyNMT*. GitHub. Retrieved 27/04/2021, from https://github.com/UKPLab/EasyNMT

Walter, D., & Ophir, Y. (2020). Strategy Framing in News Coverage and Electoral Success: An Analysis of Topic Model Networks Approach. *Political Communication*, 1–24. https://doi.org/10.1080/10584609.2020.1858379