JOURNAL OF INFORMATION TECHNOLOGY & POLITICS
https://doi.org/10.1080/19331681.2020.1731245

Automatic translation, context, and supervised learning in comparative politics

Michael Courtney, Michael Breen, Iain McMenamin, and Gemma McNulty

ABSTRACT

This paper proves that automatic translation of multilingual newspaper documents deters neither human nor computer classification of political concepts. We show how theory-driven coding of newspaper text can be automated in several languages by monolingual researchers. Supervised machine learning is successfully applied to text in English from British, Spanish, and German sources. The paper has three main findings. First, results from human coding directly in a foreign language do not differ from coding computer-translated text. Second, humans can code translated text as well as they can code untranslated prose in their mother tongue. Third, machine learning based on translated Spanish and German training sets can reproduce human coding as accurately as a system learning from English training sets.

Introduction

Comparative political text analysis requires reliable translation of textual data. Only a minority of countries in the world use English as their primary language. Greater variation in contexts will bring conceptual variables into better focus (Przeworksi & Teune, 1970; Sartori, 1970). Undertaking quantitative studies using text from parliamentary speeches, government documents, or newspapers for language contexts with which the researchers have no familiarity, is unfeasible unless texts are, firstly, converted into a single language and, second, the translated text are meaningful. Human translation of big data, and even smaller datasets, is an unrealistic pursuit for researchers. Automated computer translation systems such as *Google Translate* are a potentially cost-effective resource. While the expected efficacy of such a tool would suffice for tourists to translate basic words and phrases, rigorous assessments of their reliability and validity are required before they should be accepted in political science. In this paper, we show that automatic translation deters neither human nor machine coding of text. Theory-driven coding can be automated in several languages by monolingual researchers. Computers can perform supervised learning tasks with high accuracy regardless of whether the training data originated in English or automatically translated from Spanish or German. We contribute to the literature on automatic text analysis in political science (Eggers & Spirling, 2011; Grimmer & Stewart, 2013; Hillard, Purpura, & Wilkerson, 2007; Laver, Benoit, & Garry, 2003; Schwarz, Traber, & Benoit, 2017; Slapin & Proksch, 2008), machine learning in multilingual contexts (Kleinnijenhuis, Ruigrok, & Schlobach, 2008; Ruedin, 2013), and automatic translation applied to political text classification (De Vries, Schoonvelde, & Schumacher, 2017; Lucas et al., 2015).

Automated text analysis has emphasized induction. This unsupervised approach allows the machine to "learn" the structure of a dataset without input from humans. A popular tool is the structural topic model (Roberts et al., 2014), in which humans tell the computer the number, but not the content, of categories present in the data. The previous literature on machine-translated text for social science has focused on its application to inductive methods (De Vries et al., 2017; Lucas et al., 2015). A deductive approach to text classification of translated text, as presented in this paper, is potentially more challenging. Both approaches share the "bag-of-words" assumption, where the reliance on grammar and syntax is removed. But with the deductive approach, the computer has to

Q1 **CONTACT** Michael Breen michael.breen@dcu.ie School of Law and Government, Dublin City University, Glasnevin Dublin 9
Q2 Supplemental data for this article can be accessed on the publisher's website.

Q4

reproduce the coding of humans who need grammar and syntax to communicate, and implicitly use these elements of language to interpret the meaning and emphasis of documents. Moreover, if humans are going to train a computer to classify theoretically, then those coders will have to overcome their reliance on grammar and syntax when reading automatically translated text. The ultimate benefit of validating the efficacy of automatically translated text for supervised learning is to allow researchers more freedom to direct the focus of the computer to theoretically motivated comparative politics coding schemes.

In this paper, we present an end-to-end test of supervised learning based on multilingual texts. The analysis demonstrates that humans can code documents reliably regardless of whether coders work with English text or Spanish and German *Google*-translated to English, or where one coder performs the task using text in Spanish while others work with Spanish text translated to English. Next, the analysis shows little or no deviations in computer accuracy when the training data consist of text in English or automatically translated from Spanish or German. In some cases, accuracy is actually higher for the translated text than text originating in English.

Data

The data are randomly selected paragraphs[1] of newspaper text from the Financial Times (FT), El País (EP) and Die Welt (DW) from all sections of the newspapers[2] excluding "Sport." A team of three coded the training and test sets for the computer classification system. Inter-coder reliability scores are based on 300 paragraphs drawn equally from the three sources. The scores exceed minimum social science norms regardless of whether the text originated in English, was translated to English from Spanish or German, or coded in the original Spanish[3]. The team then coded a completed dataset of 4,485 paragraphs to perform computer accuracy tests. Paragraphs originating in Spanish and English were translated using the Google Translate API in Python.

We test the efficacy of translated text for supervised machine learning with theoretically relevant categories in English, Spanish, and German. We construct independent binary class classifiers[4] to distinguish: 1. Macroeconomic policy from everything else, 2. Microeconomic policy from everything else, 3. Political Competition from everything else, and 4. Other Policy from everything else. These categories are drawn from the Policy Agendas project (Alexandrova, Carammia, & Timmermans, 2014; Baumgartner, Green-Pedersen, & Jones, 2006). The original scheme identifies 22 policy areas[5]. We followed their general definition of policy news as requiring some explicit reference to policy-makers or advocacy of policy change. We also worked with their codebooks, which provide many examples of macro- and micro-economic policies. The definition of political competition is the same as the media studies literature on election coverage (De Vreese, Esser, & Hopmann, 2017; Dunaway & Lawrence, 2015; Lawrence, 2000; McMenamin, Flynn, O'Malley, & Rafter, 2013). It refers to the "game" of politics, including topics such as the polls, leadership, and personality, and strategy and tactics. Any other policy-related news, such as environmental policy, justice, and home affairs or foreign policy was coded as "Other Policy." We only coded policy news, so news about macroeconomic performance is not included. The next section presents the results of reliability tests of human coding.

Manual coding

Supervised learning must meet three basic conditions in order to be meaningful. First, the categories must be valid measures of theoretical concepts. Second, humans should be able to reliably reproduce each other's coding. Third, the computer should be able to reliably reproduce the humans' coding (Hillard et al., 2007; Manning, Schütze, & Raghavan, 2008). Krippendorff's α (Krippendorff, 2013) is an industry-standard measure of inter-coder reliability. While an α of above 0.8 is regarded as highly reliable, results above 0.67 are usable (Krippendorff, 2013, p. 325). Our aim was to meet the higher standard. We employ a mix of English and translated text to build multi-context politics and policy classifiers. Machine classification necessitates that all text must be standardized

to a single language for efficient identification of word features associated with each category. In total, our analysis is based on 4,485 paragraphs equally distributed between EP, DW, and the FT, 300 of which are used to assess inter-coder reliability. However, our categories are not evenly distributed in the real world. While this is not a problem for human coding, it is a potential problem at the automatic coding stage because supervised machine learning works best with balanced training sets. The distribution of our categories is outlined in Table 1.

In order to code training data efficiently, multiple coders must work concurrently and independently after demonstrating a consistent interpretation of categories. In the first instance, we must establish that Anglophone coders can reliably recognize the theoretical concepts in the three sources[6]. Table 2 presents the α scores for each of our tests. In all cases the minimum α value for usability, 0.67, specified by Krippendorff (2013) is exceeded. The first column of Table 2 reports an inter-coder reliability α score of 0.75. When all team members code text translated from Spanish, the score is 0.73. When all members code text translated from German the score is 0.79. When two coders read text translated from Spanish and one coder reads the same text, albeit the original Spanish version, the score is 0.7. These results indicate that Google-translated text can be reliably coded by humans and is appropriate for use as input to machine-learning systems. If humans can reliably code translated text, where the consistency of grammar and syntax may vary substantially, it can be expected that machines, which put no emphasis on these aspects of language, should also be able to reproduce the human codes.

Automatic coding

The computer classifiers were developed using Support Vector Machines (SVM), a popular algorithm for binary classification in social and computer science. We constructed separate "category *i*" classifiers for each category of interest; macroeconomic policy, microeconomic policy, other policy, and political competition. In each classifier, the computer attempts to distinguish between category *i* and paragraphs derived from all remaining categories and non-relevant paragraphs, or "other." We use the area-under-the-curve (AUC) statistic to find optimal cut-points that balance true positive and false-positive classification, to ultimately evaluate performance vis-à-vis the categories[7]. Higher values of AUC indicate better classifier performance. Eighty percent of category *i* paragraphs drawn from our manually coded sample were used as training data and twenty percent as test data. In the training set, half of the paragraphs are from the pool coded as category *i* and half as not category *i*. The fifty percent that are not category *i* are randomly selected from a larger pool of paragraphs reserved for feature selection and training. This distribution also applies to the test set[8]. All results presented here are five-fold cross-validated. We built classifiers using the same programming procedure

Table 1. Distribution of hand-coded paragraphs by topic.

	Policy/Politics				
	Macroeconomic policy	Microeconomic policy	Political competition	Other policy	Other news
Count	266	231	360	795	2813
Proportion	0.06	0.06	0.08	0.18	0.63

Counts and proportions are based on paragraphs sourced from the *Financial Times* and text translated to English from *El País* and *Die Welt*.

Table 2. Inter-coder reliability tests.

Language context	Original English	Translated Spanish	Translated German	Original Spanish and translated Spanish
Krippendorff's α	0.75	0.73	0.79	0.7
N	99	100	102	100

Coefficients are Krippendorff's *α* results of inter-coder agreement with three coders on randomly generated paragraphs from the Financial Times (Original English), El País (Spanish), and Die Welt (German). The coding schemes for each test are; *α1* = 1) Macroeconomic Policy, 2) Other Economic Policy, 3) Other Policy, 4) Political Competition, 5) Other.

Table 3. Variation in samples.

Samples	FT	Non-English	Mixed
Contexts	1	2	3
Proportion of translated text	0	1	0.63

seeded by humanly coded training sets that are similarly sized and reach similar levels of reliability. Therefore, variations in the performance of the classifiers should reflect differences in the texts themselves, linguistic and/or contextual. Table 3 summarizes the differences between these samples[9]. If translation is a substantial obstacle to automatic classification, the best results should be obtained for the FT, followed by a sample mixing paragraphs from all three papers, with the two translated sources displaying the weakest classifier performance. If a mixture of political and journalistic context is an obstacle to successful classification, the mixed sample should perform worse than the individual samples for each newspaper.

The text themselves were prepared for classification using minimal pre-processing procedures. This involved the removal of stopwords, numbers, punctuation, excess white space, and the inclusion of bi-grams[10]. Following Ruedin (2013), words were also reduced to their stems. The resulting Document Term Matrix (DTM) was used to identify the word features that best discriminate between the document labels by running an initial SVM on the training documents only, with the test documents withheld from this stage. The SVM provides word weights w for each feature. Features in the DTM for both the training and test documents were then multiplied by their discrimination weights. In each classifier, features associated with the category of interest, category i, have positive values, while features associated with the "other" label have negative values.

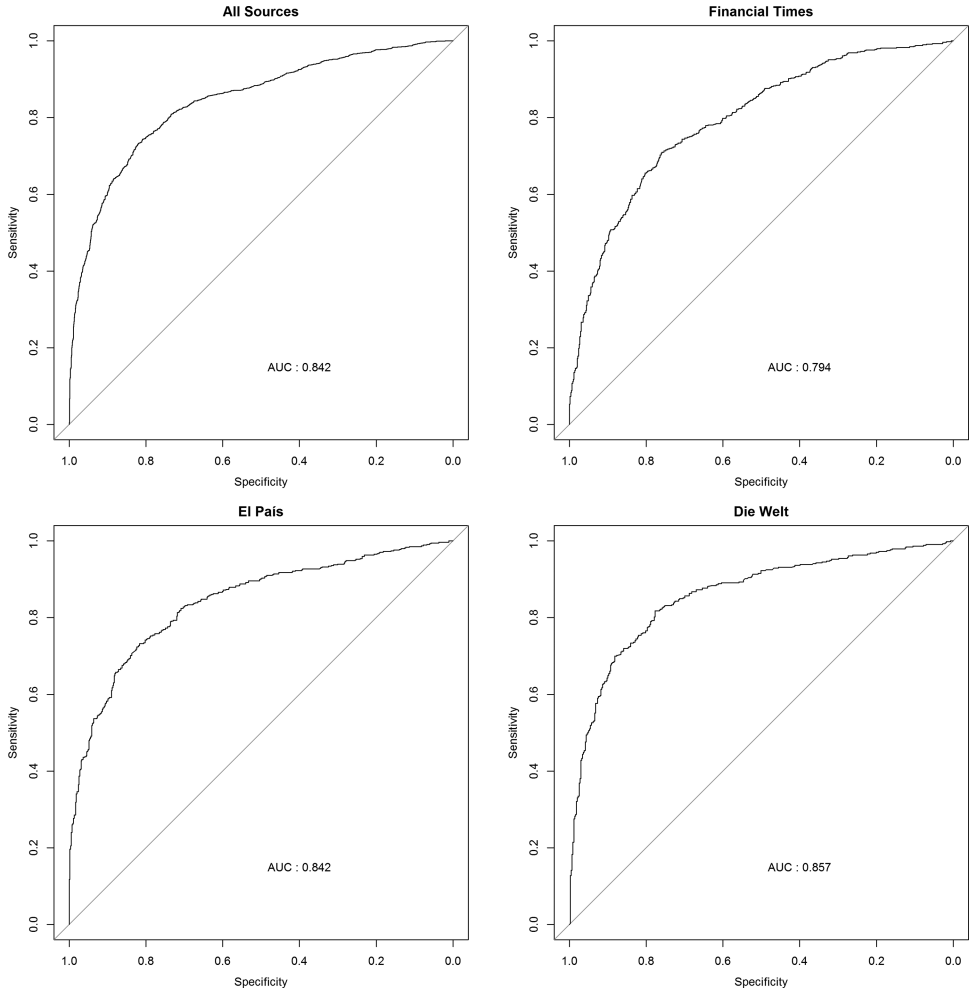
Assessment of classifier performance is conducted on each individual newspaper followed by a score with all the available training data combined. Table 4 presents the average result[11] across all classifiers of machine-learning tests by individual paper, and when all text sources, in English and translated English, are used to train the models. We expect a better result for the FT, given that English is the original language of the text. The numbers in the FT row represent the average accuracy across all our topic classifiers trained and tested exclusively on FT data. Remarkably, the AUC for the FT (0.76) is lower than those for the EP (0.8) and DW (0.81). The AUC for a mixture of sources is 0.8, where the cross-sectional data are sampled to roughly the same number of paragraphs as the individual papers. Tripling the numbers by using every paragraph we coded marginally increases the performance to 0.81. All of the scores are more than sufficient by machine-learning standards. The results show that translation is not a problem for automatic classification. They also demonstrate that a mixture of contexts is not a problem either.

Table 4. Average SVM topic scores by paper.

	AUC	Accuracy	Precision	Recall	*N*
Financial Times	0.76	0.73	0.74	0.73	1600
El País	0.80	0.77	0.77	0.79	1396
Die Welt	0.81	0.78	0.79	0.79	1400
Mixed (Sampled)	0.80	0.77	0.78	0.75	1495
Mixed	0.81	0.76	0.76	0.77	4485

To further evaluate these systems, the area-under-the-receiver-operating-characteristic (AUROC) curves for the classifiers, which indicates the trade-off between true and false positives for various levels of classification cutoffs, are plotted in Figures 1–3. In Figure 1 the focus is on the classifier's performance differentiating between all policy and politics news[12], and everything else. This is the "easy" test because the distributions of these categories are relatively evenly matched in the real data, so the training and test sets are the largest of all tests presented in this paper. The plot lines represent classifiers trained from all sources and the individual sources. The illustrations and a comparison of the AUC statistics printed in the plots show that the classifier using only FT text performs slightly worse than the Spanish, German, and mixed classifiers. This is evident by the fact that the AUROC line in the FT plot increases on the x axis (specificity, the classifier's false-positive rate) more quickly than the AUROC lines in the other plots.

Figure 2 plots the AUROC curves for classifiers focused on macroeconomic policy. Recall that macroeconomic policy paragraphs are rare in the source data. This in turn means less available training data because we use balanced training and test sets, and so we expect lower performance



Q8 **Figure 1.**

scores. Here, FT is again the weakest classifier. There are no drastic distinctions between these results and those presented in Figure 1, despite much less training data available to the macroeconomic policy classifiers.

In Figure 3 we present similar analyses for the category of political competition, which given its emphasis on personalities, is most likely to suffer from contextual over-fitting. This means that the features might be context-, rather than concept-, specific, such as a reliance on proper names like "Thatcher" or "Merkel." Therefore, the classifier would be of little use outside of the specific context for which it was trained, such as newspapers from an out-of-sample country or the same country outside of the time-period sampled. The AUC statistics here are better than the policy/politics classifier in Figure 1. Once again the computer has the greatest difficulty with the FT.

Having presented results across samples, we now present them across categories. Table 5 lists the average results by topic for classifiers trained separately from each paper, akin to averaging the ROC curves from each of the individual source classifier plots in the figures above. The range of the AUC scores is from 0.70 (Microeconomic Policy) to 0.89 (Political Competition). There are no major deviations between precision and recall for any of the topics. This means that the systems are good at identifying paragraphs coded as category *i* without falsely classifying "other" paragraphs as category *i*. Table 6 shows that the performance results are consistent with those trained from a much larger training set which pools labeled documents from the three papers, akin to the "all sources" plot lines from the figures above. The range is from 0.70 (Micro Policy and Other Policy) to 0.88 (Political Competition).

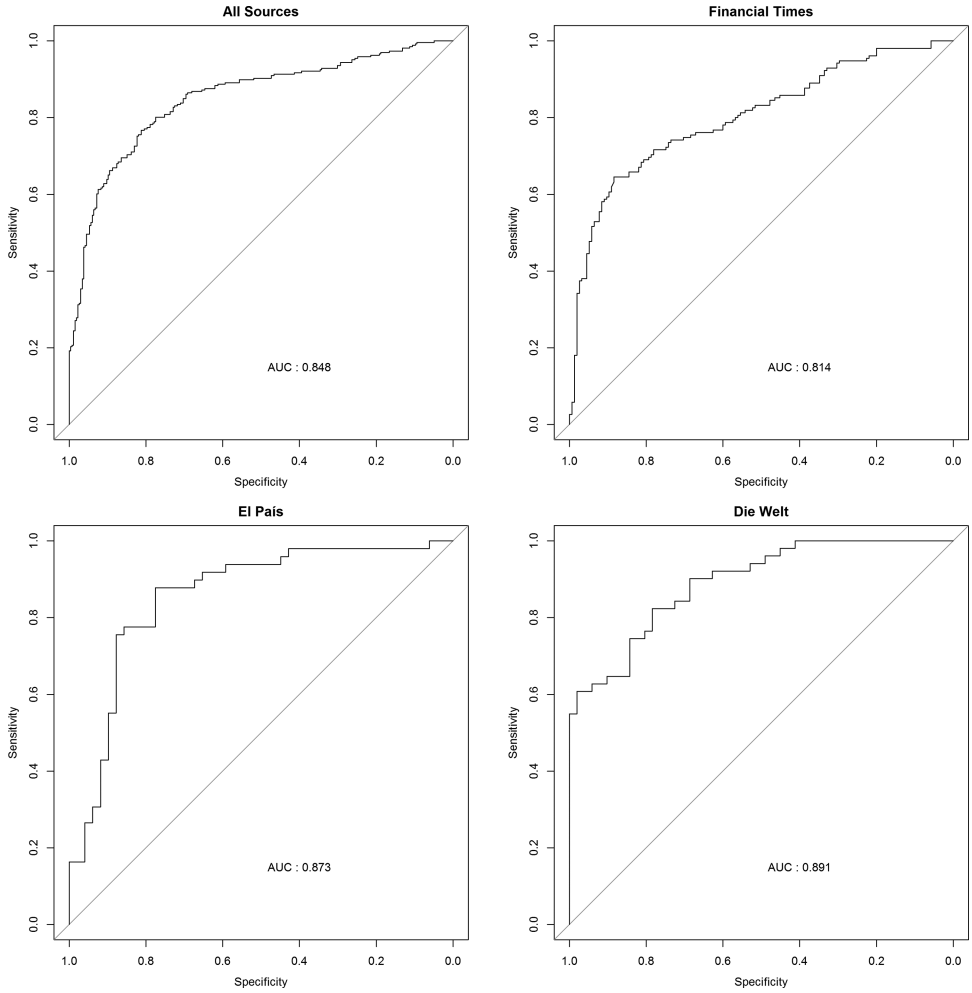


Figure 2.

The clearest pattern in our results is the absence of big differences across samples. This is very encouraging for multilingual and multi-contextual supervised learning. Nonetheless, while there is little to choose between the German, Spanish, and mixed samples, the FT is lagging somewhat on many measures. If the FT was going to be distinctive, we would have expected its performance to be superior, given the absence of translation at the pre-processing stage, and the restriction to one context. We probe why this might be by looking inside the classifiers. We wonder whether the relative weakness of the FT classifier can be explained by the consequences of humans coding in their mother tongue and in a more familiar political and economic context. Coding was holistic. Coders were instructed to code like humans, not like computers. They were discouraged from consciously seeking keywords linked to the concepts underlying the categories. After all, computers are better at finding word patterns than humans. However, when faced with a Google-translated jumble, perhaps humans could not avoid implicitly relying on keywords. These words were then used to train the computer and, because they are theoretical rather than contextual, did not reflect the idiosyncrasies of the training set and performed better in correctly classifying the test set. This implies that there are more words in the feature set for the FT that are merely statistically, rather than theoretically, associated with the categories in the training set and have higher association with non-category *i* paragraphs in the test set. However, the true explanation must be more subtle as we could not find systematic evidence of inappropriate features being associated with a weaker classifier performance[13].

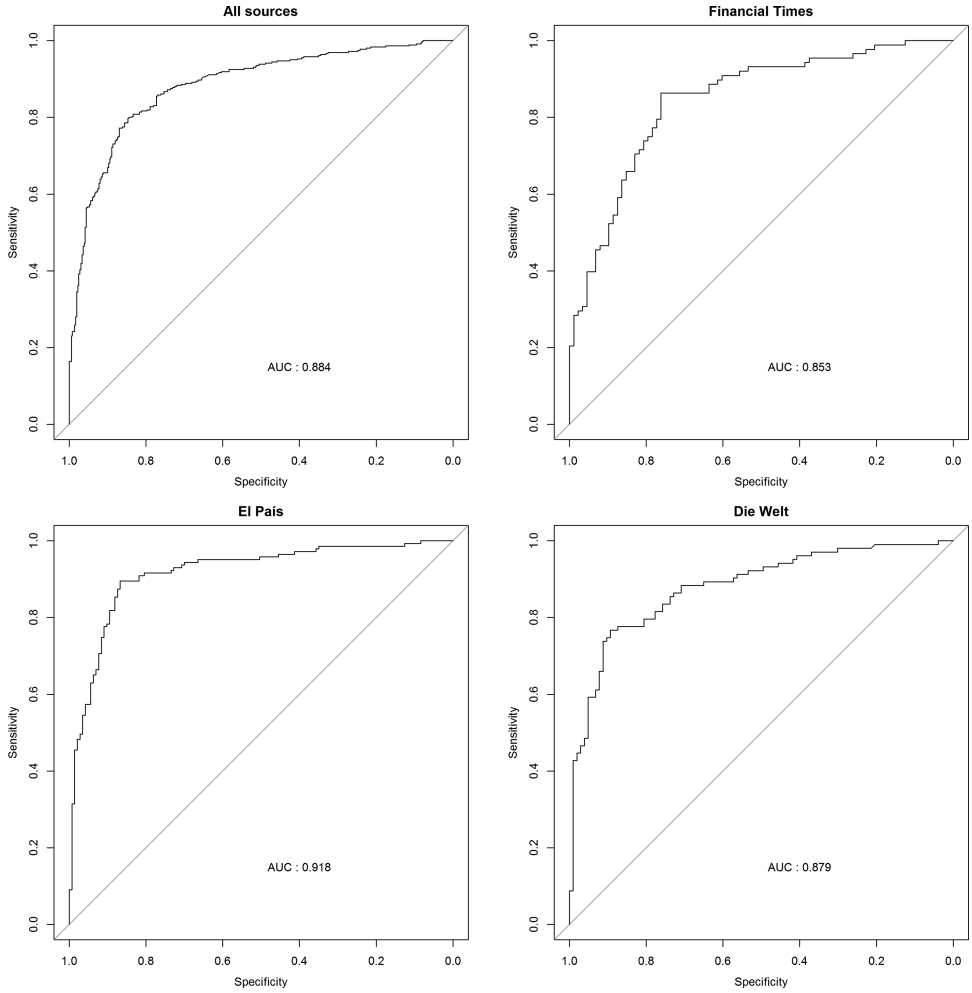


Figure 3.

Table 5. Mean SVM topic scores across independent newspaper classifiers.

	AUC	Accuracy	Precision	Recall
Policy/Politics	0.83	0.78	0.78	0.77
Policy	0.77	0.73	0.74	0.73
Political competition	0.88	0.85	0.85	0.86
Other policy	0.71	0.69	0.69	0.71
Economic policy	0.79	0.76	0.76	0.76
Macro policy	0.86	0.84	0.85	0.83
Micro policy	0.70	0.70	0.68	0.74

Table 6. SVM topic scores with pooled training.

	AUC	Accuracy	Precision	Recall
Policy/Politics	0.84	0.78	0.79	0.77
Policy	0.79	0.74	0.73	0.75
Political competition	0.88	0.84	0.85	0.83
Other policy	0.73	0.68	0.68	0.71
Economic policy	0.83	0.78	0.78	0.77
Macro policy	0.85	0.80	0.81	0.80
Micro policy	0.73	0.71	0.70	0.74

Conclusion

Three languages in three proximate contexts did not prevent a team working in English from developing an automatic political classification system that works well in all conditions. This is good news for scholars of comparative politics. It is possible to train computers to operationalize theoretical concepts across different language texts and countries without the expense of assembling a multilingual research team, while maintaining standards of reliability, validity, and accuracy. The advent of big data has already allowed social scientists to mine gigantic data sources and discover surprising patterns. Our work shows that deductively inclined comparative social scientists can test theories using big data. Applications include studies

of elite political speech, media coverage of politics, and political discourse on social media.

Our findings build on existing research on computerized text analysis in political science and its interaction with automatic translation. We contribute three distinctive conclusions. First, humans can code Google-translated text reliably. Second, human coding from untranslated text is essentially the same as human coding from automatically translated text. Third, automatic classifiers are undeterred in reproducing human coding of automatically translated text, building on previous work which employed professional translations (Ruedin, 2013). This is because the computer begins with a very simple bag of words, stripped of the grammar and syntax that are the most challenging aspects of automatic translation. The advantage of testing automated translation in a supervised rather than unsupervised context is to limit the potential for computer-induced bias at both the coding and translation stages of the test. However, given the success of our study, and the shared "bag-of-words" assumption, automatically translated texts should work well for both supervised and unsupervised analyses (De Vries et al., 2017) and, based on Lucas et al. (2015), language contexts outside of Europe.

The caveat here is that the concepts being classified must not be too granular or specialized in order to achieve high inter-coder reliability and computer accuracy scores. We found, however, that concept granularity was a more significant problem at the human coding stage than at the computer accuracy stage, even for researchers coding in English with purely English-sourced text. This inter-coder reliability challenge arose from the coding scheme, which was quite difficult for researchers without a background in political economy. For more universally understood concepts such as sentiment, the inter-coder reliability scores should be higher, leading to better computer accuracy scores. There is little evidence that automatically translating texts causes any significant research design issues, and as automated translation systems improve, such as those provided by Google, we can only expect that they become ever more reliable and useful for researchers.

For future research, it would be interesting to see whether our results generalize to languages and text sources that have less in common historically. If not, is there a threshold at which linguistic and/or contextual diversity becomes a problem? Most intriguingly, does the relative under-performance of our untranslated sample generalize and, if so, what explains this phenomenon? We have shown that some linguistic and contextual diversity need not constrain the development of theoretically based classifiers in comparative politics. Automatic translation creates more opportunities than challenges for comparative researchers.

Notes

1. Conducting machine learning at the paragraph level increases the reliability of manual coding and is more conducive to automated computer coding (Le and Mikolov, 2014).
2. Paragraphs are sampled from the full range of articles accessible on LexisNexis up to 2012. The database covers the FT (1982–2012), EP (1996–2012) and DW (2000–2012).
3. Unfortunately, no member of the team read German well enough for a similar test on direct coding from German. Neither could a German-speaker we hired from outside the research group be sufficiently trained to code the topics in English in order to perform the multilingual test.
4. The focus is on binary classifiers rather than one multi-class classifier as the classes are highly imbalanced in the real world.
5. The team could not achieve sufficient inter-coder reliability on such a granular scheme. In order to perform a reasonable test of automated translation and automated coding, the classification scheme needed to maximize reliability and validity at the human coding stage, given Ruedin and Morales (2017) point that automated coding can produce erroneous estimations for complex and specific policy areas. The eventual scheme represents the most reliable, though highly aggregated, version achievable.
6. While longer texts may undoubtedly contain multiple topics, focusing on paragraphs maximizes the likelihood of the coded unit being about one topic, or one clearly dominant topic, while containing enough information to construct a training set for machine learning. Where it is likely that the content of a paragraph could be coded for two or more topics, we coded for the dominant topic of the paragraph. We do not weight the coding toward any section of the paragraph.
7. This facilitates the extraction of probability values for class association for use in further analyses.
8. While it would be ideal to evaluate how the classifiers perform when faced with the real distribution of paragraphs, the training data are dominated by nonpolitical paragraphs, to the extent that

macroeconomic and microeconomic policy paragraphs each constitute only 5 percent of the total dataset, respectively. Testing on the real distribution would produce low precision scores as even high accuracy for non-economic policy paragraphs would lead to severe over-classification. We consider the test as presented here as one of marginal accuracy, where we evaluate whether the computer can identify an individual paragraph being relevant to the classification category or not, rather than correctly predicting the proportion of relevant paragraphs in the source data. Nevertheless, we present the AUC, accuracy, precision, and recall scores for each of our classifiers using unbalanced training and test sets in the supplemental materials.

9. The FT is slightly over-represented as it was used for initial hand-coding tests and those paragraphs were included in the final set.
10. These pre-processing decisions were appropriate for the data. The effect of pre-processing on the relationship between documents can be tested using the R package *preText* (Denny & Spirling, 2017).
11. See supplemental materials for a complete breakdown of results for each topic/newspaper combination.
12. The binary coding scheme for policy/politics is constructed by aggregating the hand codes for each of our substantive categories; macroeconomic policy, microeconomic policy, political competition, and other policy. Therefore, we have classifier that distinguishes between anything related to policy or politics, and "other news" such as financial news, sport, and culture.
13. In the supplemental material, we present lists of features associated with each classifier and investigate whether statistically correlated words or context-specific named entities are responsible for the relative weakness of the FT.

Acknowledgments

The authors would like to thank Arthur Spirling, Kenneth Benoit, Gijs Schumacher, Martijn Schoonvelde and Ronny Patz for their comments on earlier drafts of this paper.

Funding

This research was supported by the Irish Research Council Grant Number GOIPD/2016/253.

Notes on contributors

Michael Courtney is a Statistician at the Central Statistics Office, Ireland.

Michael Breen is an Associate Professor of Politics at the School of Law and Government, Dublin City University.

Iain McMenamin is Professor of Politics at the School of Law and Government, Dublin City University.

Gemma McNulty is Research Associate at the Clinton Institute, University College Dublin.

References

Alexandrova, P., Carammia, M., & Timmermans, A. (2014). High politics: The policy agenda of the European Council, 1975-2011. In F. Foret & Y. S. Rittelmeyer (Eds.), *The European Council and European governance* (pp. 53–72). London: Routledge.

Baumgartner, F. R., Green-Pedersen, C., & Jones, B. D. (2006). Comparative studies of policy agendas. *Journal of European Public Policy*, *13*(7), 959–974. doi:10.1080/13501760600923805

De Vreese, C. H., Esser, F., & Hopmann, D. N. (2017). *Comparing political journalism*. London: Routledge.

De Vries, E., Schoonvelde, M., & Schumacher, G. (2017). *Lost in translation? Evaluating the usefulness of machine translation for bag-of-words text models*. Open Science Framework.

Denny, M. J., & Spirling, A. (2017, April 6th-10th). *Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it*. Paper presented at the 75th Annual Meeting of the Mid-West Political Science Association, Chicago, IL, USA. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2849145

Dunaway, J., & Lawrence, R. G. (2015). What predicts the game frame? Media ownership electoral context and campaign news? *Political Communication*, *32*(1), 43–60. doi:10.1080/10584609.2014.880975

Eggers, A., & Spirling, A. (2011). *Partisan convergence in executive-legislative interactions modeling debates in the House of Commons* (Unpublished manuscript).

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 1–31. doi:10.1093/pan/mps028

Hillard, D., Purpura, S., & Wilkerson, J. (2007). Computer-assisted topic classification for mixed- methods social science research. *Journal of Information Technology and Politics*, *4*(4), 31–46. doi:10.1080/19331680801975367

Kleinnijenhuis, J., Ruigrok, N., & Schlobach, S. (2008). Good news or bad news? Conducting sentiment analysis on Dutch text to distinguish between positive and negative relations. *Journal of Information Technology and Politics*, *5*(1), 73–94. doi:10.1080/19331680802154145

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3rd ed.). Los Angeles: Sage.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, *97*(2), 311–331. doi:10.1017/S0003055403000698

Lawrence, R. G. (2000). Game-framing the issues: Tracking the strategy frame in public policy news. *Political Communication*, *17*(2), 93–114. doi:10.1080/105846000198422

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. doi:10.1093/pan/mpu019

Manning, C. D., Schütze, H., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

McMenamin, I., Flynn, R., O'Malley, E., & Rafter, K. (2013). Commercialisation and election framing. *The International Journal of Press/Politics*, *18*(2), 433–448.

Przeworksi, A., & Teune, H. (1970). *The logic of comparative social inquiry*. New York, NY: Wiley Interscience.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Albertson, B., … Rand, D. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082. doi:10.1111/ajps.12103

Ruedin, D. (2013). The role of language in the automatic coding of political texts. *Swiss Political Science Review*, *19*(4), 539–545. doi:10.1111/spsr.12050

Ruedin, D., & Morales, L. (2017). Estimating party positions on immigration: Assessing the reliability and validity of different methods. *Party Politics*. doi:10.1177/1354068817713122

Sartori, G. (1970). Concept misinformation in comparative politics. *American Political Science Review*, *64*(4), 323–344. doi:10.2307/1958356

Schwarz, D., Traber, D., & Benoit, K. (2017). Estimating intra-party preferences: Comparing speeches to votes. *Political Science Research and Methods*, *5*(2), 379–396. doi:10.1017/psrm.2015.77

Slapin, J. B., & Proksch, S. O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, *52*(3), 254–277. doi:10.1111/j.1540-5907.2008.00338.x