Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings

Hauke Licht

Cologne Center for Comparative Politics, Institute of Political Science and European Affairs, University of Cologne, Cologne , Germany. E-mail: hauke.licht@wiso.uni-koeln.de

Abstract

Established approaches to analyze multilingual text corpora require either a duplication of analysts' efforts or high-quality machine translation (MT). In this paper, I argue that multilingual sentence embedding (MSE) is an attractive alternative approach to language-independent text representation. To support this argument, I evaluate MSE for cross-lingual supervised text classification. Specifically, I assess how reliably MSE-based classifiers detect manifesto sentences' topics and positions compared to classifiers trained using bag-of-words representations of machine-translated texts, and how this depends on the amount of training data. These analyses show that when training data are relatively scarce (e.g., 20K or less-labeled sentences), MSE-based classifiers can be more reliable and are at least no less reliable than their MT-based counterparts. Furthermore, I examine how reliable MSE-based classifiers label sentences written in languages *not* in the training data, focusing on the task of discriminating sentences that discuss the issue of immigration from those that do not. This analysis shows that compared to the within-language classification benchmark, such "cross-lingual transfer" tends to result in fewer reliability losses when relying on the MSE instead of the MT approach. This study thus presents an important addition to the cross-lingual text analysis toolkit.

Keywords: multilingual embedding, multilingual text analysis, supervised machine learning

1 Introduction

Text-as-data methods allow generating insights from text corpora that could otherwise be analyzed only by investing large amounts of human effort, time, and financial resources (cf. Grimmer and Stewart 2013). However, when applied in cross-lingual research, many existing quantitative text analysis methods face limitations (Baden *et al.* 2021), such as picking up on language differences instead of substantively more interesting patterns (cf. Lind *et al.* 2021a). Analyzing multilingual corpora as-is, in turn, necessitates analysts to duplicate their efforts (Lucas *et al.* 2015; Reber 2019).

Machine translation (MT) has been proposed and validated as a remedy to these limitations (e.g., Lucas *et al.* 2015). However, when relying on commercial MT services, translating large multilingual corpora can be expensive. Translating only dictionary keywords or the words retained after tokenizing documents in their original languages (e.g., Proksch *et al.* 2019; Reber 2019), in turn, can lead to incorrect translations.

This paper presents an alternative approach to cross-lingual quantitative text analysis. Instead of translating texts, they are represented in a language-independent vector space by processing them through a pre-trained multilingual sentence embedding (MSE) model. Existing pre-trained models enable semantically meaningful text representation and are publicly available for replicable and resource-efficient use in research. However, only a minority of the texts used to pre-train these models stem from the political domain.

To do so, I focus on cross-lingual text classification as an application. First, I rely on a dataset compiled by Düpont and Rachuj (2022) that records machine-translated and original sentences of election manifestos in the *Comparative Manifestos Project* (CMP) corpus (Volkens *et al.* 2020). I assess how reliable MSE-based classifiers perform in classifying sentences' topics and positions

Political Analysis (2023)
vol. 31: 366–379
DOI: 10.1017/pan.2022.29

Published
26 January 2023

Corresponding author
Hauke Licht

Edited by
Jeff Gill

© The Author(s), 2023. Published by Cambridge University Press on behalf of the Society for Political Methodology.

https://doi.org/10.1017/pan.2022.29 Published online by Cambridge University Press

Table 1. Sentences in multilingual example corpus.

	Language	Text
doc_1	English	"We will fight unemployment."
doc_2	German	"Wir werden die Arbeitslosigkeit reduzieren."

compared to classifiers trained using bag-of-words (BoW) representations of machine-translated texts (the "MT+BoW" benchmark). I also include MT-based classifiers in this comparison that rely on translations by the open-source M2M model (Fan *et al.* 2021) to compare between "free" alternatives. This analysis shows that relying on MSEs for text representation enables training classifiers that are no less reliable than their MT-based counterparts. Moreover, I find that relying on free MT (i.e., the M2M model) instead of Goggle's commercial MT service reduces the reliability of classifiers only slightly.

Next, I examine how these classifiers' reliability depends on the amount of labeled data available for training. This analysis shows that adopting the MSE approach tends to result in more reliable cross-lingual classifiers than the MT+BoW approach and, at least, likely results in no less reliable classifiers—particularly when working with training data sizes typically available in applied research. However, as more training data are added, this comparative advantage decreases.

Lastly, I compare how MT+BoW and MSE-based classifiers perform in *cross-lingual transfer classification*, that is, classifying sentences written in languages *not* in the training data. Annotated text corpora are often limited in their country coverage, and extending them to new countries beyond their original language coverage is a promising application of cross-lingual classifiers. I probe the MSE and MT+BoW approaches in this task based on a dataset compiled by Lehmann and Zobel (2018) covering eight languages that records human codings of manifesto quasi-sentences into those discussing the immigration issues and those that do not. Specifically, I conduct an extensive text classification experiment to estimate how much less reliable cross-lingual transfer classification is compared to the "within-language" classification benchmark examined in the first two analyses. This experiment shows that cross-lingual transfer tends to result in fewer reliability losses when relying on the MSE instead of the MT approach.

2 Approaches to Cross-Lingual Quantitative Text Analysis

The goal of quantitative text analysis is to infer indicators of latent concepts from text (Grimmer and Stewart 2013; Laver, Benoit, and Garry 2003). Achieving this goal in multilingual applications is challenging, because similar ideas are express with different words in different languages. The two sentences in Table 1 illustrate this. In both, authors pledge to lower unemployment, but these ideas are expressed in different words in English and German.

Hence, the goal of *cross-lingual* quantitative text analysis is to obtain identical measurements for documents if they indicate the same concept independent from the language they are written in (cf. Lucas *et al.* 2015, 258). There are currently two dominant approaches to tackle this challenge: "separate analysis" and "input alignment" through MT.

2.1 Established Approaches

The first approach is to separately analyze documents in their original languages. For example, in the case of human coding, separate analysis requires human coders to annotate each language-specific subcorpus. Analysts can then use these annotations to directly estimate quantities of interest or to train language-specific supervised text classifiers.[1] A significant shortcoming of the

1 Similarly, separate analysis requires adapting keywords to each target language (cf. Lind *et al.* 2019; Proksch *et al.* 2019) or align estimated topics across languages (cf. Lind *et al.* 2021a).

https://doi.org/10.1017/pan.2022.29 Published online by Cambridge University Press

Table 2. Representations of sentences in Table 1 after multilingual sentence embedding. Rows report sentences' d-dimensional embedding vectors; columns report embedding dimensions.

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	$\cdots$	e_{d-1}	e_d
doc_1	0.335	0.909	0.412	0.044	0.764	0.750	0.800	0.885	$\cdots$	0.449	0.488
doc_2	0.379	0.870	0.400	0.056	0.771	0.738	0.839	0.841	$\cdots$	0.423	0.449

Note: These data serve illustrative purposes only.

separate analysis approach thus is that analysts need to duplicate their efforts for each language present in a corpus (Lucas *et al.* 2015; Reber 2019). This duplication makes separate analysis a relatively resource-intensive strategy.

An alternative approach is *input alignment*. The idea of input alignment is to represent documents in a language-independent way that enables analysts to apply standard quantitative text analysis methods to their multilingual corpora instead of analyzing them language by language (Lucas *et al.* 2015). Translating text inputs into one target language using commercial MT services, such as *Google Translate*, has been established as a best practice to achieve this. Specifically, with the *full-text translation* approach, texts are translated as-is into the "target" language (e.g., English).[2] Full-text translated documents can then be pre-processed and tokenized into words and phrases (n-gram tokens) to obtain monolingual BoW representations of originally multilingual documents.[2] This approach has been shown to enable reliable dictionary analysis (Windsor, Cupit, and Windsor 2019), topic modeling (de Vries, Schoonvelde, and Schumacher 2018; Lucas *et al.* 2015; Maier *et al.* 2021; Reber 2019), and supervised text classification (Courtney *et al.* 2020; Lind *et al.* 2021b).

However, when researchers rely on commercial MT services, translating full texts can be very expensive, rendering this approach relatively resource-intensive, too (but see Lind *et al.* 2021b). An alternatives is to tokenize documents in their original languages, and only translate the resulting language-specific sets of words and phrases (e.g., Düpont and Rachuj 2022; Lucas *et al.* 2015). Similar to the full-text translation approach, *token translation* enables representing documents as BoW vectors in the target language.[2] Moreover, it is relatively resource-efficient because it implies translating fewer characters. However, token translation implies translating words and phrases outside their textual contexts, which can result in incorrect translations that impair the quality of BoW text representations.

Hence, researchers' dependence on commercial MT services for full-text translation has created a trade-off between cost efficiency and text representation quality. The recent publication of pre-trained MT models by (for example, M2M by Fan *et al.* 2021) promises to break this dependence, and I evaluate this possibility in my analyses below. However, I first present *MSE* as an alternative, MT-free approach to language-independent text representation.

2.2 Multilingual Sentence Embedding

MSE is a method to represent sentence-like texts[3] as fixed-length, real-valued vectors such that texts with similar meaning are placed close in the joint vector space *independent* from their language. Because MSE allows representing documents written in different languages in the same feature space, it presents an alternative input alignment approach to cross-lingual quantitative text analysis. Table 2 and Figure 1 illustrate this for the two sentences in Table 1. Because these

2 For a running example, see Section A of the Supplementary Material.
3 "Sentence-like" typically includes even short paragraphs. For example, the knowledge-distilled models I evaluate below can embed sentence with up to 128 tokens, that is, between 73 and 94 words (see Figure S.3 in the Supplementary Material).

https://doi.org/10.1017/pan.2022.29 Published online by Cambridge University Press

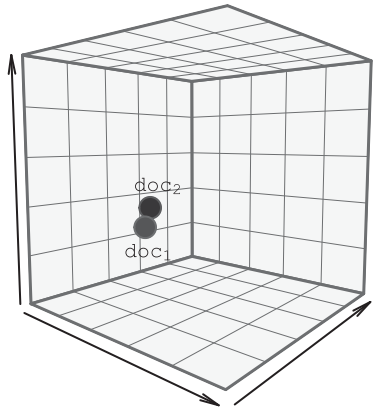

Figure 1. Schematic depiction of multilingual sentence embedding of example sentences in Table 1.
Note: Depicting embedding in three dimensions serves illustrative purposes only.

sentences are semantically very similar, their embeddings are very similar and they are hence placed close in the embedding space.

The idea of representing textual inputs as dense vectors (i.e., "embed" them) to encode their semantic relationships is old (Harris 1954). *Word embedding* models obtain such vectors for short *n*-grams (e.g., Mikolov *et al.* 2013; Pennington, Socher, and Manning 2014), and have already been popularized in the social sciences (cf. Garg *et al.* 2018; Rodman 2020; Rodriguez and Spirling 2021). *Sentence embedding* models obtain such fixed-length vectors for sentence-like texts such that semantically similar texts are placed relatively close in the embedding space (e.g., Conneau *et al.* 2017). MSE methods, in turn, obtain such vectors in a language-agnostic way.

Researchers have developed different MSE methods in recent years (e.g., Artetxe and Schwenk 2019; Reimers and Gurevych 2020; Yang *et al.* 2020). They commonly use corpora recording translations of sentences in different languages ("parallel sentences") as inputs to train a neural network model that learns to induce a sentence embedding of the input text.[4] The *Language-Agnostic Sentence Embedding Representations* (LASER) model proposed by Artetxe and Schwenk (2019), for example, trains to translate parallel sentences and learns to induce language-agnostic sentence embeddings as an intermediate step.

Once "pre-trained" on large amounts of parallel data, MSE models can be used to embed texts they have not seen during pre-training.[5] Indeed, existing pre-trained MSE models have been shown to obtain sentence embeddings that (i) encode texts' semantic similarity independent of language and (ii) provide critical signals to achieve competitive performances in a wide range of natural language processing tasks (e.g., Artetxe and Schwenk 2019; Reimers and Gurevych 2020; Yang *et al.* 2020). Moreover, publicly available MSE models have been pre-trained on large parallel corpora covering very many languages (e.g., 113 in the case of LASER; see Section B.1 of the Supplementary Material).

This suggests that MSE is an attractive alternative to the MT approach to input alignment discussed above. Instead of BoW count vectors of documents' machine-translated texts, one combines their MSE vectors in a document-feature matrix (cf. Table 2).[6] As elaborated below, this enables using MSEs as features to train cross-lingual supervised text classifiers.

4 See Section B of the Supplementary Material.
5 This is the central idea of "transfer learning" (cf. Ruder *et al.* 2019).
6 Columns of the document-feature matrix thus record embedding dimensions instead of BoW tokens.

https://doi.org/10.1017/pan.2022.29 Published online by Cambridge University Press

3 Empirical Strategy

To assess whether MSE enables reliable cross-lingual analyses of political texts, I evaluate this approach for *cross-lingual supervised text classification* (CLC).[7] The overarching goal of my analyses is to establish whether relying on pre-trained MSE models for text representation enables reliable measurement in relevant political text classification tasks. The reliability of classifiers trained using BoW representations of machine-translated texts—the "MT+BoW" approach—constitutes the reference point in this assessment.

I focus on supervised text classification as an application for three reasons. First, it figures prominently in quantitative text analysis (e.g., Barberá *et al.* 2021; Burscher, Vliegenthart, and De Vreese 2015; D'Orazio *et al.* 2014; Rudkowsky *et al.* 2018). However, in contrast to topic modeling (Chan *et al.* 2020; cf. Lind *et al.* 2021a), relatively view attention has been paid to evaluate translation-free CLC approaches besides the separate analysis strategy (but see Glavaš, Nanni, and Ponzetto 2017).

Second, MSEs can be directly integrated into the supervised text classification pipeline. Labeled documents are first sampled into training and validation data splits. Documents in the training data are then embedded using a pre-trained MSE model and their MSEs used as features to train a supervised classifier.

Third, the reliability of supervised text classifiers can be evaluated using clearly defined metrics (cf. Grimmer and Stewart 2013, 279). One first applies a classifier to predict the labels of documents in the validation data split. Comparing a classifier's predictions to "true" labels then allows quantifying its reliability in labeling held-out documents with metrics such as the F1 score.

3.1 Analysis 1: Comparative Reliability

I first assess how reliable MSE-based classifiers perform in classifying the topic and position of sentences in political parties' election manifestos compared to classifiers trained with the MT+BoW approach.[8] Classifying the topical focus and left–right orientation of political texts are among the main applications of text classification methods in comparative politics research (Benoit *et al.* 2016; Burscher *et al.* 2015; Osnabrügge *et al.* 2021; Quinn *et al.* 2010). Assessing the comparative reliability of MSE-based classifiers in these tasks is thus relevant to a large group of researchers. Moreover, the data I use are representative of other annotated political text corpora with sentence-like texts or paragraphs as units of annotation (e.g., Barberá *et al.* 2021; Baumgartner, Breunig, and Grossman 2019; Rudkowsky *et al.* 2018).

3.1.1 Data. The annotated sentences used in this analysis stem from a subset of the CMP corpus (Volkens *et al.* 2020)[9] for which machine-translated full texts are available from the replication materials of Düpont and Rachuj (2022, henceforth D&R).[10] D&R study programmatic diffusion between parties across countries and have translated a sample of manifestos covering 12 languages (see Table 3) with *Google Translate* to validate their token translation-based measurement strategy. Sentences in D&R's original data are not labeled, however, and I have hence matched them to original quasi-sentence-level CMP codings.[11] This allows training and evaluating topic and position classifiers with both the MT+BoW and MSE approaches on the *same* data and hence their direct comparison.

7 Please refer to Licht (2022a) for replication data and code.
8 The position classification task is to assign each sentence into one of the categories "left," "right," or "neutral," or into the "uncoded" category; the topic classification task is to assign each sentence into one of the CMP's seven topic categories (see Section C.1 of the Supplementary Material) or the "uncoded" category (cf. Osnabrügge, Ash, and Morelli 2021).
9 The CMP records human-annotated election manifestos of political parties from 48 developed countries.
10 I kindly thank Nils Düpont and Martin Rachuj for sharing these data with me.
11 See Section C.2 of the Supplementary Material.

https://doi.org/10.1017/pan.2022.29 Published online by Cambridge University Press

Table 3. Description of data sources combined with primary data recorded in the *Comparative Manifestos Project* corpus (Volkens *et al.* 2020).

	Analyses 1 and 2	Analysis 3
Data source	Düpont and Rachuj (2022)	Lehmann and Zobel (2018)
Task	Position classification Topic classification	Discriminating immigration/integration from other issue mentions
Unit of annotation	Sentence	Quasi-sentence
Labeled samples	70,999	222,847
Language coverage	Catalan, Danish, Dutch, Finnish, French, Galician, German, Italian, Norwegian, Portuguese, Spanish, and Swedish	Danish, Dutch, English Finnish, German, Norwegian, Spanish, and Swedish

3.1.2 Classifier Training and Evaluation. The resulting corpus records 70,999 sentences. I randomly sample these sentences *five times* into 50:50 training and validation data splits.[12] This ensures that the out-of-sample performance estimates I report are not dependent on the data split.

For each training dataset and classification task, I train *five classifiers*: two MT+BoW and three MSE-based ones. In the case of the MT+BoW approach, I train one classifier using D&R's original *Google Translate* translations and another one using translations of the same sentence I have obtained using the open-source M2M model (Fan *et al.* 2021). This allows assessing whether relying on "free" MT instead of a commercial service impairs the reliability of BoW-based classifiers. In both cases, I apply a five-times repeated fivefold cross-validation (5×5 CV) procedure to select the best-performing classifier.[13]

In the case of the MSE approach, I train classifiers relying on three different publicly available pre-trained MSE models.[13] The LASER model (Artetxe and Schwenk 2019) already discussed in Section 2.2 and two models that have been trained for sequence alignment of parallel sentences by adopting the multilingual "knowledge distillation" procedure proposed by Reimers and Gurevych (2020): a *multilingual Universal Sentence Encoder* (mUSE, Yang *et al.* 2020), and an XLM-RoBERTa (XLM-R) model (Conneau *et al.* 2020). Embedding texts with different pre-trained models allows comparing their suitability for political text classification applications.

For each training dataset and task, I then evaluate the five resulting classifiers on sentences in the corresponding validation datasets and bootstrap 50 F1 score estimates per classifier. I summarize these estimates in Figure 2 below.

3.2 Analysis 2: Comparative Effectiveness

Next, I examine how these classifiers' reliability depends on the amount of labeled data available at training time (cf. Barberá *et al.* 2021; Burscher *et al.* 2015). The amount of digitized texts available for quantitative analyses is increasing, but collecting annotations for these data is usually very resource-intensive (cf. Benoit *et al.* 2016; Hillard, Purpura, and Wilkerson 2008). As a consequence, applied researchers can often afford to collect annotations for only a few documents in their target corpus. It is thus practically relevant to know which of the text representation approaches I compare proves more reliable in data-scarce scenarios.

12 The Training datasets thus record 35,496 sentences.
13 See Section D.2 of the Supplementary Material.

https://doi.org/10.1017/pan.2022.29 Published online by Cambridge University Press

3.2.1 Data. I use the same data as in Analysis 1.

3.2.2 Classifier Training and Evaluation. The training and evaluation procedure I adopt is the same as in Analysis 1, with two exception. First, I vary the size of the training datasets from 5% to 45% (in 5 percentage point increments) of the target corpus, whereas in Analysis 1, I have trained on 50%. So the smallest (largest) training dataset in Analysis 2 records 3,549 (31,948) labeled sentences.[14] Second, I rely only on the knowledge-distilled XLM-R model for sentence embedding because it results in the most reliable MSE-based classifiers in Analysis 1.

3.3 Analysis 3: Cross-Lingual Transfer

Last, I investigate which of the two text representation approaches I compare enables more reliable cross-lingual transfer classification, that is, classifying documents written in languages *not* present in the training data. Such "out-of-language" classification is a promising application of cross-lingual text classifiers. Annotated text corpora are often limited in their country coverage. Training cross-lingual text classifiers on these data promises to extend their coverage to new countries beyond their original language coverage.

3.3.1 Data. I rely on a dataset compiled by Lehmann and Zobel (2018, henceforth L&Z) covering eight languages (see Table 3). Their data record human codings of election manifesto quasi-sentences into those that discuss the immigration issues and those that do not, and I train cross-lingual classifiers for this binary classification task.

The example of identifying passages in political documents that discuss the issue of immigration is an ideal case for probing the reliability of supervised text classifiers in cross-lingual transfer. The politicization of immigration by the radical right since the 1990s has raised scholars' interest in studying how governments, mainstream parties, and the media change their attention to this issue. However, then-existing databases, such as the CMP, lacked suitable indicators to address this question quantitatively. Despite scholars' impressive efforts to obtain such indicators by means of content analysis, the resulting annotated corpora are often limited in their geographic coverage (cf. Lehmann and Zobel 2018; Ruedin and Morales 2019). The methodological problem of expanding the coverage of these corpora by means of cross-lingual transfer classification has thus considerable practical relevance.

3.3.2 Classifier Training and Evaluation. I examine this problem in an experimental setup designed to estimate how much less reliable cross-lingual transfer classification is compared to the "within-language" classification benchmark examined in Analyses 1 and 2. The basic idea of this setup is to use quasi-sentences written in some "source languages" to train a classifier that is then evaluated on held-out quasi-sentences. Repeated for many different combinations of source languages, I can estimate how reliably a given set of held-out quasi-sentences can be classified when the languages they are written in are among the source languages ("within language" classification) compared to when they are not ("out of language" classification, i.e., cross-lingual transfer).[15]

4 Results

4.1 Comparative Reliability

Figure 2 reports the reliability of position and topic classifiers in terms of their cross-class mean F1 scores. Comparing average cross-class mean F1 scores shows that the best MSE-based classifier (the one trained using XLM-R embeddings) outperforms the benchmark classifier (the one relying on commercial MT) in topic classification while performing as reliably in position classification.

14 See Section D.1 of the Supplementary Material.
15 See Section D.3 of the Supplementary Material for a detailed description of the experimental setup.

https://doi.org/10.1017/pan.2022.29 Published online by Cambridge University Press

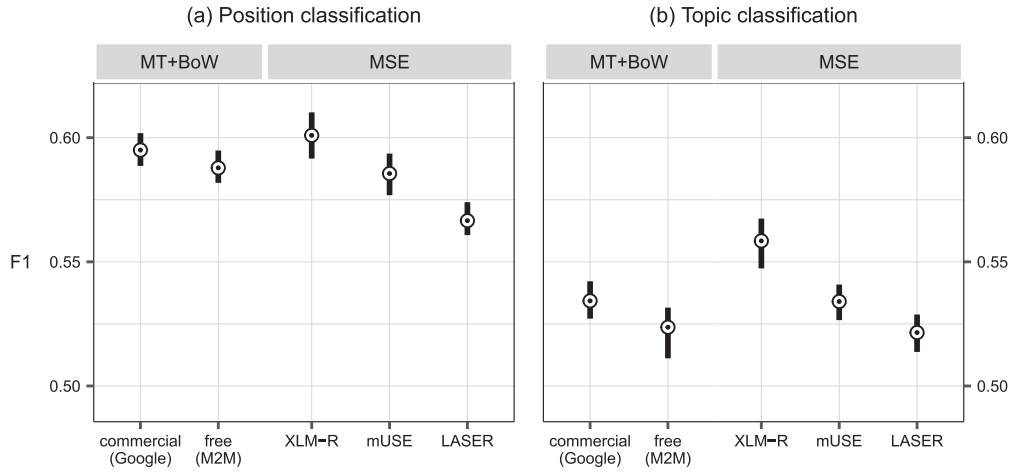

Figure 2. Cross-class mean F1 scores of classifiers trained using different text representation approaches: bag-of-words obtained from machine-translated texts (MT+BoW) and multilingual sentence embeddings obtained from original texts (MSE). Panel (a) reports results for classifying manifesto sentences' positions; panel (b) for classifying their topic. Data plotted summarize 50 bootstrapped cross-class mean F1 scores (excluding the `uncoded` category) for five classifiers per task and approach. Points are averages of bootstrapped estimates, and vertical lines span the 95% most frequent values.

In addition, the best MSE-based classifier is more reliable than the BoW-based classifier relying on free MT in topic classification and a similar tendency can be observed in the case of position classification. There is thus *no* indication that training using MSEs instead of BoW representations of machine-translated texts substantially reduces the reliability of cross-lingual text classifiers.[16]

Note, however, that absolute F1 scores indicate that the reliability achieved with either approach is rather modest. One reason for this may be the strong class imbalance across label categories.[17] The poor quality of human annotations in the CMP corpus is another likely reason (cf. Mikhaylov, Laver, and Benoit 2012). As shown in Section 4.3, better performance can be achieved with less noisy labels.

Comparing MSE-based classifiers, it is notable that using the knowledge-distilled XLM-R model for sentence embedding results in the most reliable classifiers in both tasks, whereas using LASER consistently results in the least reliable classifiers (cf. Reimers and Gurevych 2020). With regard to differences in MT-based classifiers' F1 scores, it is striking that the classifiers relying on translation with the open-source M2M model label held-out sentences only slightly less reliably than those relying on Google's commercial MT service.[18]

Finally, when comparing classifiers' reliability across languages (see Figure S.7 in the Supplementary Material), it is notable that the F1 scores of the classifiers relying on commercial MT and the classifiers trained using XLM-R embeddings are strongly correlated for both tasks.[19] Moreover, the standard deviation of languagewise differences in classifiers' F1 scores is modest in both tasks[20] and these differences are mostly indistinguishable from zero (accounting for variability in bootstrapped F1 scores; see Figure S.8 in the Supplementary Material). Furthermore, with 0.15, the correlation in language-specific F1 scores between tasks is rather low, suggesting that task-specific factors contribute significantly to between-language differences in classifiers' reliability. These findings are reassuring since they provide little evidence of systematic language bias in the pre-trained embedding model.[21]

16 This holds for precision and recall (see Figure S.6 in the Supplementary Material).
17 Predictive performance is lowest for minority label classes in both tasks (see Table S.15 in the Supplementary Material).
18 The difference is 0.72 [0.04, 1.31] F1 points for position and 1.06 [0.42, 1.80] points for topic classification.
19 0.66 [0.64, 0.68] for position and 0.75 [0.74, 0.77] for topic classification.
20 0.02 for position and 0.02 for topic classification.
21 Language bias would mean that the pre-trained model exhibits poorer representation quality for some languages.

https://doi.org/10.1017/pan.2022.29 Published online by Cambridge University Press

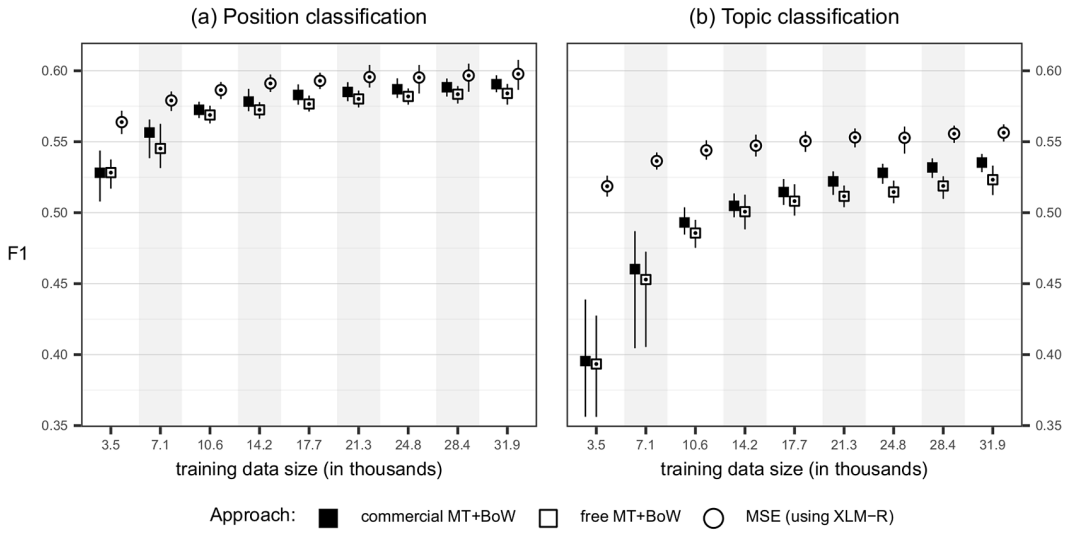

Figure 3. Cross-class mean F1 scores of classifiers trained using different text representation approaches as a function of training data size. Panel (a) reports results for classifying manifesto sentences' positions; panel (b) for classifying their topic. Data plotted summarize 50 bootstrapped estimates of the cross-class average F1 scores (excluding the `uncoded` category) of five classifiers per training data size, approach, and task. Points are averages of bootstrapped estimate, and vertical lines span the 95% most frequent values.

In summary, Figure 2 provides evidence that the MSE approach enables training position and topic classifiers that are no less reliable than classifiers trained using BoW representations of machine-translated texts. Moreover, relying on a free MT model instead of a commercial service reduces the reliability of BoW classifiers only slightly in the two tasks examined here.

4.2 Comparative Effectiveness as a Function of Training Data Size

However, how effective using MSEs for text representation is compared to the MT approach also depends on the amount of labeled data available at training time. This is shown in Figure 3 by plotting the F1 scores of classifiers trained on different amounts of labeled data when adopting these different text representation approaches.

Three patterns stand out from the data presented in Figure 3. First, MSE-based classifiers tend to outperform their MT-based counterparts when training data are scarce. Second, as more training data are added, this comparative advantage decreases. Third, relying on the open-source M2M model for MT instead of Google's commercial service consistently results in less reliable classifiers, but, in line with the findings presented above, differences in terms of F1 score points are overall very small.

The first pattern is more pronounced in the case of topic classification. Taking variability in bootstrapped F1 estimates into account, the MSE-based topic classifiers outperform the ones relying on commercial MT across the entire range of training data sizes examined here.[22] What is more, when training on only 3.5K labeled sentences, the topic classifiers relying on MT are only slightly more reliable than human coders (Mikhaylov *et al.* 2012, 85), whereas MSE-based classifiers perform relatively well.

While this comparative advantage is less pronounced in the case of position classification,[23] the BoW-based classifiers outperform their MSE-based counterparts at *none* of the training size values examine here. This suggests that the amount of labeled data needed to reach a "tipping point" at which MT-based classifiers begin outperforming their MSE-based counterparts is quiet

22 I use the 97.5% percentile of differences in bootstrapped F1 scores as the criterion.

23 The position classifiers trained using MSEs outperform the ones trained relying on commercial (free) MT "only" if training on 21.3K (31.9K) labeled sentences (or less).

https://doi.org/10.1017/pan.2022.29 Published online by Cambridge University Press

large and likely larger than what is typically available in applied political and communication science research.

Nevertheless, adding more training data results in greater F1 improvements for MT-based than for MSE-based classifiers in both tasks. As a consequence, the comparative reliability advantage of the MSE approach tends to decrease. This difference between approaches in how adding more training data affects classifiers' reliability is not surprising. Training on BoW representations, classifiers learn to identify tokens in the training data that allow reliable classification. The features enabling reliable classification based on BoW representations are thus "domain-specific." In contrast, classifiers trained using MSEs hinge on the representations the embedding model has learned to induce while pre-training on corpora that overwhelmingly stem from other domains. Learning domain-specific BoW features from machine-translated texts should thus eventually trump the "transfer learning" logic underpinning the MSE approach.[24] But as emphasized above, the amount of labeled data needed to reach this "tipping point" is likely very large.

This reasoning also helps explaining why the range of training data sizes at which MSE-based classifiers are more reliable than their MT-based counterparts is larger for topic than for position classification. Identifying tokens that reliably predict held-out sentences topical focus among seven different policy issue areas from strongly imbalanced training data is more difficult than identifying tokens that discriminate between three positional categories. This gives the MSE-based classifiers a greater head start in topic classification.

Viewed together, the data presented in Figure 3 suggest that adopting the MSE approach tends to enable more—but at least no less—reliable cross-lingual classification than training on BoW representations of machine-translated texts. While the comparative reliability advantage of the MSE approach decreases as the training data size increases, none of the training data sizes examined here results in BoW-based classifiers that outperform their MSE-based counterpart. This suggests that the MSE approach is particularly suited when working with training data sizes typically available in applied research.

4.3 Cross-Lingual Transfer Classification

But how do the MSE and MT+BoW approaches to cross-lingual text classification perform when applied to label documents written in languages not present in the training data? As described above, I rely on the L&Z dataset recording codings of manifestos' quasi-sentences into those that discuss the immigration issue and those that discuss other issues to address this question.

To establish a baseline estimate of the two approaches' reliability in this binary classification task, I have first trained MSE- and MT-based classifiers on a balanced dataset recording a total of 10,394 quasi-sentences sampled from all eight languages and evaluated them on held-out quasi-sentences.[25] In this within-language classification setup, both approaches result in very reliable classifiers. With average F1 scores of 0.85 [0.84, 0.86] and 0.83 [0.82, 0.85], respectively, the MSE-based and MT-based classifiers are about equally reliable. Given that they were trained on about 10K labeled quasi-sentences, this finding is in line with the results presented in Figure 3. However, the immigration issue classifiers are much more reliable. Moreover, their language-specific average F1 scores are all above 0.79 in the case of the MSE-based classifier and above 0.75 in the case of the MT-based classifier. This suggest that both approaches should enable training classifiers on the L&Z data that perform well in cross-lingual transfer.

However, Figure 4 provides evidence that cross-lingual transfer tends to result in less F1 reductions (relative to the within-language classification benchmark) when relying on the MSE

24 That the BoW-based classifiers tend to catch up with their MSE-based counterparts as more training data are added supports this expectation.

25 I have again used the knowledge-distilled XLM-R model for sentence embedding and the M2M model for machine translation.

https://doi.org/10.1017/pan.2022.29 Published online by Cambridge University Press

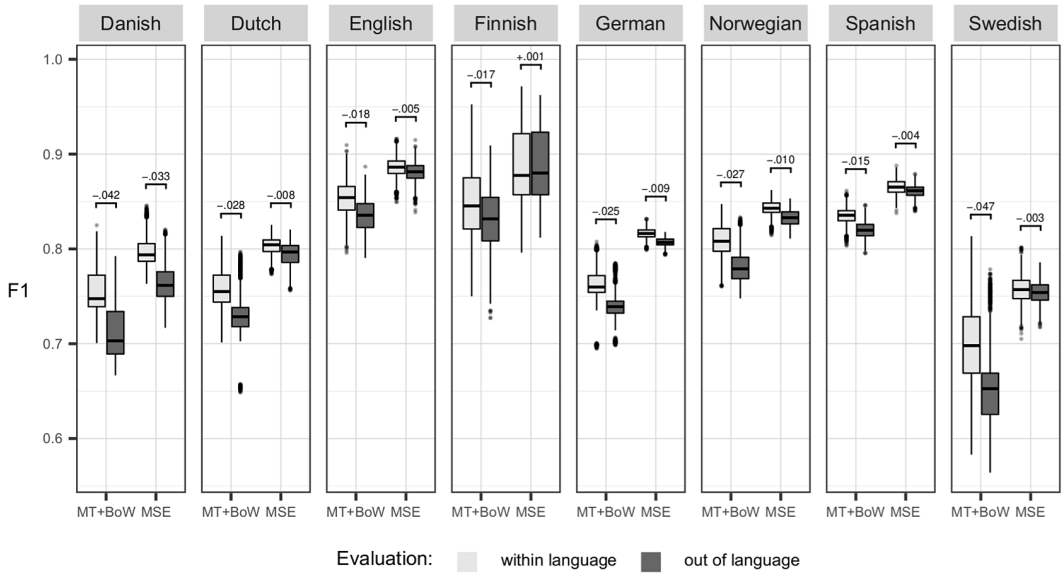

Figure 4. Results of cross-lingual transfer classification experiments conducted on the Lehmann and Zobel (2018) data. Box plots summarize F1 scores achieved by classifiers when evaluated on held-out quasi-sentences written in the language reported in the header of panel columns (the "target" language) by approach (*x*-axis values) and by whether the target language has been in the training data ("within language" evaluation) or not ("out of language" evaluation, i.e., cross-lingual transfer). Differences between the mean F1 score achieved in within-language (resp. out-of-language) evaluation (printed above box plot pairs) estimate the "reliability cost" of cross-lingual transfer.

instead of the MT+BoW approach. For example, the "reliability cost"[26] of cross-lingual transfer into Danish is about 2.8 F1 score points with the MT+BoW approach and only 0.8 F1 score points with the MSE approach. This pattern is consistent across languages recorded in the L&Z data. Moreover, the average F1 scores achieved by MSE-based classifiers when predicting quasi-sentences written in languages that were not in the training data (i.e., *out-of-language evaluation*) are higher than those of their MT-based counterparts. This suggests that the MSE approach enables more reliable cross-lingual transfer classification.

5 Conclusion

In this paper, I have argued that relying on MSEs presents an attractive alternative approach to text representation in cross-lingual analysis. Instead of translating texts written in different languages, they are represented in a language-independent vector space by processing them through a pre-trained MSE model.

To support this claim empirically, I have evaluated whether relying on pre-trained MSE models enables reliable cross-lingual measurement in supervised text classification applications. Based on a subset of the CMP corpus (Volkens *et al.* 2020) for which machine-translated full texts are available, I have first assessed how reliably MSE-based classifiers perform in classifying manifesto sentences' topics and positions compared to classifiers trained using BoW representations of machine-translated texts. Moreover, I have evaluated how these classifiers' reliability depends on the amount of labeled data available for training. These analyses show that adopting the MSE approach tends to result in more reliable cross-lingual classifiers than the MT+BoW approach and, at least, likely results in no less reliable classifiers. However, as more training data are added, this comparative advantage decreases. Moreover, I show that relying on an open-source MT model (Fan *et al.* 2021) reduces MT-based classifiers' reliability only slightly.

26 I quantifying the "reliability cost" of cross-lingual transfer for each approach and target language by subtracting the mean F1 score achieved in within-language evaluation from that achieved in out-of-language evaluation.

https://doi.org/10.1017/pan.2022.29 Published online by Cambridge University Press

In addition, I have assessed how MSE- and MT-based classifiers perform when applied to classify sentences written in a language that was not present in their training data (i.e., cross-lingual transfer). Using a dataset compiled by Lehmann and Zobel (2018) that records human codings of manifesto quasi-sentences into those discussing the immigration issues and those that do not, I show that cross-lingual transfer tends to result in fewer reliability losses when relying on the MSE instead of the MT approach compared to the within-language classification benchmark examined in the first two analyses.

These results suggest that MSE is an important addition to applied researchers' text analysis toolkit, especially when their resources to collect labeled data are limited. When they want to train a cross-lingual classifier on a small to modestly sized labeled corpus, adopting the MSE approach can benefit the reliability of their classifier but, at least, will likely not harm it. Moreover, when the country coverage of their labeled corpus is limited and extending it by means of cross-lingual transfer would require "out-of-language" classification, my analyses suggest that adopting the MSE instead of the MT approach should result in fewer additional classification error.

Acknowledgments

I thank Tarik Abou-Chadi, Elliott Ash, Pablo Barberá, Nicolai Berk, Theresa Gessler, Fabrizio Gilardi, Benjamin Guinaudeau, Christopher Klamm, Fabienne Lind, Stefan Müller, Sven-Oliver Proksch, Martijn Schoonvelde, Ronja Sczepanski, Jon Slapin, Lukas Stötzer, and three anonymous reviewers for their thoughtful comments on this manuscript. I acknowledge support by the *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1-390838866.

Data Availability Statement

Replication code for this article has been published in Code Ocean, a computational reproducibility platform that enables users to run the code and can be viewed interactively at https://doi.org/10.24433/CO.5199179.v1 (Licht 2022a). A preservation copy of the same code and data can also be accessed via Dataverse at https://doi.org/10.7910/DVN/OLRTXA (Licht 2022b).

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2022.29.

References

Artetxe, M., and H. Schwenk. 2019. "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond." *Transactions of the Association for Computational Linguistics* 7: 597–610. https://doi.org/10.1162/tacl_a_00288.

Baden, C., C. Pipal, M. Schoonvelde, and M. A. C. G. van der Velden. 2021. "Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda." *Communication Methods and Measures* 16 (1): 1–8. https://doi.org/10.1080/19312458.2021.2015574.

Barberá, P., A. E. Boydstun, S. Linn, R. McMahon, and J. Nagler. 2021. "Automated Text Classification of News Articles: A Practical Guide." *Political Analysis* 29 (1): 19–42. https://doi.org/10.1017/pan.2020.8.

Baumgartner, F. R., C. Breunig, and E. Grossman (eds.). 2019. *Comparative Policy Agendas: Theory, Tools, Data*. Oxford:Oxford University Press.

Benoit, K., D. Conway, B. E. Lauderdale, M. Laver, and S. Mikhaylov. 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.

Burscher, B., R. Vliegenthart, and C. H. De Vreese. 2015. "Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts?" *The Annals of the American Academy of Political and Social Science* 659 (1): 122–131. https://doi.org/10.1177/0002716215569441.

Chan, C.-H., et al. 2020. "Reproducible Extraction of Cross-Lingual Topics (rectr)." *Communication Methods and Measures* 14 (4): 285–305. https://doi.org/10.1080/19312458.2020.1812555.

https://doi.org/10.1017/pan.2022.29 Published online by Cambridge University Press

Conneau, A., et al. 2020. "Unsupervised Cross-Lingual Representation Learning at Scale." In Jurafsky, D., J. Chai, N. Schluter, and J. Tetreault (eds.). *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747.

Conneau, A., D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. 2017. "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data." In Palmer, M., R. Hwa, S. Riedel (eds.). *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 670–680. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1070.

Courtney, M., M. Breen, I. McMenamin, and G. McNulty. 2020. "Automatic Translation, Context, and Supervised Learning in Comparative Politics." *Journal of Information Technology & Politics* 17 (3): 208–217. https://doi.org/10.1080/19331681.2020.1731245.

D'Orazio, V., S. T. Landis, G. Palmer, and P. Schrodt. 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.

De Vries, E., M. Schoonvelde, and G. Schumacher. 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.

Düpont, N., and M. Rachuj. 2022. "The Ties That Bind: Text Similarities and Conditional Diffusion among Parties." *British Journal of Political Science* 52 (2): 613–630. https://doi.org/10.1017/S0007123420000617.

Fan, A., et al. 2021. "Beyond English-Centric Multilingual Machine Translation." *Journal of Machine Learning Research* 22 (107): 1–48.

Garg, N., L. Schiebinger, D. Jurafsky, and J. Zou. 2018. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." *Proceedings of the National Academy of Sciences of the United States of America* 115 (16): E3635–E3644. https://doi.org/10.1073/pnas.1720347115.

Glavaš, G., F. Nanni, and S. P. Ponzetto. 2017. "Cross-Lingual Classification of Topics in Political Texts." In Hovy, D., S. Volkova, D. Bamman, D. Jurgens, B. O'Connor, O. Tsur, A. S. Doğruöz (eds.). *Proceedings of the Second Workshop on NLP and Computational Social Science*, 42–46. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-2906.

Grimmer, J., and B. M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297. https://doi.org/10.1093/pan/mps028.

Harris, Z. S. 1954. "Distributional Structure." *WORD* 10 (2–3): 146–162. https://doi.org/10.1080/00437956.1954.11659520.

Hillard, D., S. Purpura, and J. Wilkerson. 2008. "Computer-Assisted Topic Classification for Mixed-Methods Social Science Research." *Journal of Information Technology & Politics* 4 (4): 31–46. https://doi.org/10.1080/19331680801975367.

Laver, M., K. Benoit, and J. Garry. 2003. "Extracting Policy Positions from Political Texts using Words as Data." *The American Political Science Review* 97 (2): 311–331.

Lehmann, P., and M. Zobel. 2018. "Positions and Saliency of Immigration in Party Manifestos: A Novel Dataset Using Crowd Coding." *European Journal of Political Research* 57 (4): 1056–1083.

Licht, H. 2022a. "Replication Data for: Cross-Lingual Classification of Political Texts using Multilingual Sentence Embeddings." Code Ocean V1. https://doi.org/10.24433/CO.5199179.v1.

Licht, H. 2022b. "Replication Data for: Cross-Lingual Classification of Political Texts using Multilingual Sentence Embeddings." Harvard Dataverse V1. https://doi.org/10.7910/DVN/OLRTXA.

Lind, F., J.-M. Eberl, O. Eisele, T. Heidenreich, S. Galyga, and H. G. Boomgaarden. 2021a. "Building the Bridge: Topic Modeling for Comparative Research." *Communication Methods and Measures* 16: 96–114. https://doi.org/10.1080/19312458.2021.1965973.

Lind, F., J.-M. Eberl, T. Heidenreich, and H. G. Boomgaarden. 2019. "When the Journey Is as Important as the Goal: A Roadmap to Multilingual Dictionary Construction." *International Journal of Communication* 13: 21.

Lind, F., T. Heidenreich, C. Kralj, and H. G. Boomgaarden. 2021b. "Greasing the Wheels for Comparative Communication Research: Supervised Text Classification for Multilingual Corpora." *Computational Communication Research* 3 (3): 1–30. https://doi.org/10.5117/CCR2021.3.001.LIND.

Lucas, C., R. A. Nielsen, M. E. Roberts, B. M. Stewart, A. Storer, and D. Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23 (2): 254–277. https://doi.org/10.1093/pan/mpu019.

Maier, D., C. Baden, D. Stoltenberg, M. De Vries-Kedem, and A. Waldherr. 2021. "Machine Translation vs. Multilingual Dictionaries Assessing Two Strategies for the Topic Modeling of Multilingual Text Collections." *Communication Methods and Measures* 16: 19–38. https://doi.org/10.1080/19312458.2021.1955845.

Mikhaylov, S., M. Laver, and K. R. Benoit. 2012. "Coder Reliability and Misclassification in the Human Coding of Party Manifestos." *Political Analysis* 20 (1): 78–91. https://doi.org/10.1093/pan/mpr047.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." In Burges, C.J., L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.). *Advances in Neural Information Processing Systems*, vol. 26. Curran Associates, Inc.

Osnabrügge, M., E. Ash, and M. Morelli. 2021. "Cross-Domain Topic Classification for Political Texts." *Political Analysis* First view: 1–22. https://doi.org/10.1017/pan.2021.37.

https://doi.org/10.1017/pan.2022.29 Published online by Cambridge University Press

Pennington, J., R. Socher, and C. Manning. 2014. "GloVe: Global Vectors for Word Representation." In Moschitti, A., B. Pang, W. Daelemans (eds.). *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162.

Proksch, S.-O., W. Lowe, J. Wäckerle, and S. Soroka. 2019. "Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches." *Legislative Studies Quarterly* 44 (1): 97–131. https://doi.org/10.1111/lsq.12218.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54 (1): 209–228. https://doi.org/10.1111/j.1540-5907.2009.00427.x.

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.

Reimers, N., and I. Gurevych. 2020. "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation." In Webber, B., T. Cohn, Y. He, Y. Liu (eds.). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 4512–4525. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.365.

Rodman, E. 2020. "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors." *Political Analysis* 28 (1): 87–111. https://doi.org/10.1017/pan.2019.23.

Rodriguez, P. L., and A. Spirling. 2021. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." *The Journal of Politics* 84 (1): 101–115. https://doi.org/10.1086/715162.

Ruder, S., M. E. Peters, S. Swayamdipta, and T. Wolf. 2019. "Transfer Learning in Natural Language Processing." In Sarkar, A., and M. Strube (eds.). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 15–18. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-5004.

Rudkowsky, E., M. Haselmayer, M. Wastian, M. Jenny, Š. Emrich, and M. Sedlmair. 2018. "More than Bags of Words: Sentiment Analysis with Word Embeddings." *Communication Methods and Measures* 12 (2–3): 140–157. https://doi.org/10.1080/19312458.2018.1455817.

Ruedin, D., and L. Morales. 2019. "Estimating Party Positions on Immigration: Assessing the Reliability and Validity of Different Methods." *Party Politics* 25 (3): 303–314. https://doi.org/10.1177/1354068817713122.

Volkens, A., et al. 2020. The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2020a. Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB). https://doi.org/10.25522/manifesto.mpds.2020a.

Windsor, L. C., J. G. Cupit, and A. J. Windsor. 2019. "Automated Content Analysis across Six Languages." *PLoS One* 14 (11): e0224425. https://doi.org/10.1371/journal.pone.0224425.

Yang, Y., et al. 2020. "Multilingual Universal Sentence Encoder for Semantic Retrieval." In Celikyilmaz, A., T.-H. Wen (eds.). *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 87–94. https://doi.org/10.18653/v1/2020.acl-demos.12.

https://doi.org/10.1017/pan.2022.29 Published online by Cambridge University Press