Political Science Research and Methods (2026), 1–19
doi:10.1017/psrm.2026.10102

ORIGINAL ARTICLE

Validating open-source machine translation for quantitative text analysis

Hauke Licht[1], Ronja Ida Sczepanski[2], Moritz Laurer[3,4] and Ayjeren Bekmuratovna[5]

[1]Department of Political Science and Digital Science Center, University of Innsbruck, Innsbruck, Austria; [2]Centre for European Studies and Comparative Politics, Sciences Po, Paris, France; [3]Cohere Inc., Toronto, ON, Canada; [4]Communication Science, Vrije Universiteit Amsterdam, Amsterdam, Netherlands and [5]Independent researchers
Corresponding author: Hauke Licht; Email: hauke.licht@uibk.ac.at

(Received 1 April 2024; revised 24 February 2025; accepted 21 June 2025)

Abstract
Machine translation (MT) is an essential tool in many multilingual computational text analysis applications. However, relying on commercial services like Google Translate or DeepL limits reproducibility and can be expensive. This paper assesses the viability of a reproducible, transparent, and affordable alternative: open-source MT models. We ask whether using open-source MT models instead of commercial services substantially changes the measurements obtained from multilingual corpora by extending an influential study by de Vries et al. and contributing an original study focusing on Transformer-based supervised text classification. Our findings reveal negligible differences in results between the two MT approaches, suggesting that open-source MT models are highly valuable tools for multilingual text analysis.

Keywords: machine translation; multilingual topic modeling; multilingual transformers; reproducibility

1. Introduction

Political scientists often want to study phenomena in text materials such as political speeches, administrative documents, or news reports written in different languages. Machine translation (MT) is a popular strategy for researchers who want to apply quantitative text analysis methods to such multilingual text collections (e.g., Baum and Zhukov, 2019; Dancygier and Margalit, 2020; Düpont and Rachuj, 2022; Barberá *et al.*, 2024; cf. Baden *et al.*, 2022; Dolinksy *et al.*, 2022; Licht and Lind, 2023). MT allows bridging language barriers by transferring documents written in different languages into a single target language. This not only enables researchers to analyze the resulting monolingual documents with standard text-as-data methods (e.g., Lucas *et al.*, 2015; Vries *et al.*, 2018; Reber, 2019; Windsor *et al.*, 2019; Courtney *et al.*, 2020; Lind *et al.*, 2021). It can also facilitate cross-lingual measurements' validation and increase studies' transparency (Baden *et al.*, 2022; Licht and Lind, 2023).

However, many scholars rely on commercial services for MT, such as Google Translate or DeepL (but see Laurer *et al.*, 2023; Licht, 2023; Mate *et al.*, 2023). This approach comes with clear limitations. First, using commercial services limits transparency and reproducibility because the underlying translation models are closed-sourced (Chan *et al.*, 2020). Second, machine-translating large amounts of text can be expensive because commercial services charge users for each translated character.[1]

[1]See Table 1 on page 6.

© The Author(s), 2026. Published by Cambridge University Press on behalf of EPS Academic Ltd. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

Table 1. Comparison of costs in Euro and compute time arising when translating a fixed amount of text with commercial services or the M2M open-source MT model

N characters	API/model	GPU	Costs (EUR)	Run time (h)
18 mio.	Google Trans.		326.93	-/- [a]
18 mio.	DeepL		359.26	-/- [a]
18 mio.	M2M-418 m	A100	2.36	1.6
18 mio.	M2M-418 m	V100	2.36	3.9[b]
18 mio.	M2M-418 m	T4	2.36	10.4[b]

[a]We disregard the time elapsed for sending API requests.
[b]Estimates based on relative efficiency relative to A100 GPU.

This paper argues for the viability of a transparent, reproducible, and affordable alternative: using open-source models for MT. Open-source MT models, such as OPUS-MT (Tiedemann and Thottingal, 2020) or Facebook Research's M2M model (Fan *et al.*, 2021), facilitate the reproducibility and transparency of research because they are versioned and publicly available for download. Moreover, existing models cover many translation directions similar to Google Translate or DeepL (see Table A2). And last but not least, using open-source models for MTs allows researchers to translate large text corpora without paying fees to a commercial translation service.

While open-source MT models promise reproducible and cheaper MT, applied researchers must know whether these models enable reliable cross-lingual quantitative text analysis of political text corpora. We thus assess whether machine-translating multilingual corpora with open-source models instead of a commercial service (Google Translate or DeepL) reduces reliability and yields substantially different measurements. We focus on two central computational text analysis methods: bag-of-words Latent Dirichlet Allocation (LDA) topic modeling and transformer (Blei *et al.*, 2003) model fine-tuning for supervised text classification (cf. Wankmüller, 2024). First, we extend the seminal study by de Vries *et al.* (2018), who evaluated the MT approach for cross-lingual topic modeling. Second, we present an original large-scale comparative study of the reliability of MT for supervised text classification with transformer-based language models. The corpora used in these studies cover several Indo-European languages as well as languages from the Uralic and Turkic language families.

Our findings support the conclusion that open-source MT models can be a reliable replacement for commercial services when applying bag-of-words and transformer-based text analysis methods. We find only minor differences between the measurements obtained from corpora translated with open-source models and commercial services. In our topic modeling study, we find that the topics estimated by a model fitted to parliamentary speeches we machine-translated with the open-source M2M model are as similar to the topic model fitted to human-translated speeches as those estimated by a comparable model fitted to speech translations generated with Google Translate. In addition, both machine-translation-based topic models allocate speeches to similar topics as the human-translation-based model. In our supervised text classification study, we find that transformer-based text classifiers fine-tuned using translations from open-source MT models perform, on average, only 0.009 F1 score points worse than comparable classifiers fine-tuned using translations from a commercial MT model as input. Importantly, this result holds across many non-European languages in our sample.

We conclude that open-source MT is a very valuable addition to researchers' multilingual text analysis toolkit. Our study adds to a growing body of work on multilingual text analysis methods and has direct practical implications for applied researchers. To facilitate the wider adoption of free MT in applied research, we provide an online translation application.[2]

[2]The application is available online via Google Colab.

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

2. Machine translation for quantitative text analysis

MT has been extensively validated for various political text analysis tasks and languages. While a comprehensive review of this literature is beyond the scope of this article (see Licht and Lind, 2023), Table A1 in the Supplementary Material provides an overview of this literature, and we highlight key insights from this literature below.

In their seminal study, Lucas *et al.* (2015) argue that comparative researchers can use MT to translate multilingual corpora into English to enable their joint analysis with standard bag-of-words methods. They demonstrate this strategy by analyzing Arabic and Chinese social media posts.

The study by de Vries *et al.* (2018) was the first to support the argument of Lucas et al. (2015) with more extensive comparative evidence. They base their study on a subset of the Europarl parallel corpus (Koehn, 2005), which contains the original text of speeches held in the European Parliament and their translations into the official languages of the EU by its expert translators. The authors constructed several bilingual parallel corpora from this dataset by pairing English texts' expert translations with their German, Spanish, French, and Polish versions. De Vries *et al.* (2018) demonstrate that MT with *Google Translate* enables translation of texts from German, Spanish, French, Danish, and Polish into English with sufficient reliability for bag-of-words topic modeling when compared to the benchmark of translation by human experts.

The study by De Vries *et al.* (2018) has been highly influential. It is frequently cited in applied research to justify an MT approach to cross-lingual bag-of-words text analyses (e.g., Barberá *et al.*, 2024). In addition, it has been the point of departure for several other methodological advancements. For example, Reber (2019) systematically compares the reliability of alternative translation strategies and commercial MT services. Düpont and Rachuj (2022), in turn, evaluate the MT approach for comparing the textual similarity of documents across languages. Courtney *et al.* (2020) present evidence on the reliability of MT for bag-of-words supervised text classification. They examine whether supervised text classifiers trained on English-language MTs of originally Spanish or German texts classify held-out texts as accurately as classifiers trained on English texts. More recently, Mate *et al.* (2023) have made a first step in adding to this finding by examining how the translation of Polish and Hungarian parliamentary speeches affects the reliability of transformer-based classifiers (see also Laurer *et al.*, 2023).

2.1. Open-source MT: a reproducible, transparent, and affordable alternative

The results summarized above underscore that MT can enable reliable and valid multilingual text analysis. However, using commercial services such as Google Translate raises concerns about the reproducibility and transparency of research (Chan *et al.*, 2020) and can be relatively expensive.

We argue that open-source MT models, such as OPUS-MT (Tiedemann and Thottingal, 2020) and M2M (Fan *et al.*, 2021), offer a reproducible, transparent, and cost-effective alternative.[3] Using open-source MT models is more transparent than a commercial service and ensures the reproducibility of analyses because researchers can freely download models and document the specific model version they used in their study. This accessibility further enables running MT on local servers or user-grade hardware, which is often necessary to conform to data privacy requirements, for example, when analyzing interview data or survey responses. Furthermore, freely available open-source MT models match many of the several hundred translation directions enabled by Google Translate and other commercial providers. For example, OPUS-MT and M2M can translate 151 and 99 languages to English, respectively (see Table A2 in the Supplementary Material). Last but not least, using open-source models is cheaper because researchers with some programming experience only need to invest in GPU computing resources[4] instead of paying a translation fee to a commercial service.

[3]Please refer to the Supplementary Material, Section A.1, for details of these MT models.

[4]GPUs are "Graphics Processing Units" designed for efficient parallel processing.

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

2.1.1. Transparency and reproducibility

Using commercial MT services, researchers have no control over the specific version of the model used for translation, as they are "closed source." This prevents others from replicating their results later because the MT system used originally may have changed in the meantime (Chan *et al.*, 2020). This problem does not arise when using open-source MT models. They are typically versioned and available from publicly accessible platforms or repositories, such as the Hugging Face Model hub.[5] Researchers can therefore document the exact version of the model they have used, making research using open-source MT models reproducible.

Second, open-source MT models are transparent. The research teams providing them typically document the parallel corpora and model architecture used to train their models. Furthermore, it is a well-established best practice to report model performance in predefined test sets (benchmarks). This enables researchers to make informed decisions about which MT model to use in a specific application (cf. Licht and Lind, 2023), for example, by assessing the reliability of the available models in the languages they want to translate.[6] In contrast, the information available about the performance of commercial services' MT models often does not meet scientific standards, nor is it transparent what data has been used for training.

2.1.2. Cost efficiency and resource requirements

Using commercial services to machine-translate large text corpora can be expensive, as one pays for each translated character. For example, Google charges users of their *Cloud Translation* service U.S. $ 20 per 1 million characters, often implying three to four-digit sums for the translation.[7]

Accordingly, researchers' reliance on commercial services creates an undesirable barrier for those with limited budgets (Baden *et al.*, 2022). Open-source MT models can lower this barrier. They are freely available for download and use, and researchers thus do not have to pay translation fees. Instead, the only financial cost arising when using open-source MT models results from the energy consumed for computing and, if necessary, from using cloud servers. Overall, this cost efficiency makes open-source MT models an attractive option for researchers on a tight budget and may even help level the playing field for smaller research teams.

Table 1 provides a cost comparison between two popular commercial MT services (Google Translate and DeepL) and a popular open-source MT model (M2M 418 M). The two main costs that MT incurs are money and time. For commercial translation services (APIs), financial costs are calculated on a per-character basis. For open-source models, the costs are based on GPU costs, which are easily accessible through services like *Google Colab*.[8] Our cost estimates show that, financially, using open-source models is significantly cheaper than commercial services. Translating 18 million characters costs more than EUR 300 via an API and less than EUR 3 with an open-source model.

However, using open-source MT models requires more computation time and expertise. Especially on older GPUs, translations with open-source models can take many hours to complete. Moreover, running the required software on a GPU requires some additional expertise (e.g., moderate Python programming skills). To tilt the balance further in favor of the open-source approach, we

[5] See https://huggingface.co/models?pipeline_tag=translation.

[6] See Table A2 in the Supplementary Material for an overview of models and services language coverage.

[7] Translating the non-English corpora in the ParlSpeech2 dataset (Rauh and Schwalbach, 2020) would, for example, cost U.S. $ 12986.56 for the Danish corpus (649 million characters) and U.S. $ 17756.08 for the Dutch corpus (887 million characters).

[8] For Google Translate, the costs are U.S. $ 20/1 million characters (as of March 2024). For DeepL, these costs are U.S. $ 20/1 million characters plus a required monthly baseline fee of U.S. $ 5 for the *Pro* account (as of March 2024). Google Colab makes GPUs accessible either for free (with lower reliability) or for around EUR 11 for a fixed budget of compute hours. We used a set of 500 long parliamentary speeches for our cost estimate, which amounted to 18 million characters. We then computed the API costs or the costs for different GPUs accessible via Google Colab Pro. The costs are derived empirically for the A100 GPU and then estimated for the other two types of GPUs.

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

thus provide an interactive Google Colab-based online translation application, code base, and tutorial.[9] We have designed our app to ease the use of open-source MT models for researchers with no Python programming skills. Our accompanying tutorial and code base provide researchers with basic Python programming experience with a template that they can adapt to their needs. Given that the level of technical prowess in the (computational) social science community is steadily increasing (cf. Baden *et al.*, 2022), these resources should lower the above-discussed barriers to using open-source MT models.

2.1.3. Advantage over generative LLMs

Although not a point of comparison in this paper, we find it worth pointing out that open-source MT models also compare favorably in terms of cost and resource efficiency to another alternative: using generative large language models (LLMs) for translation. LLMs such as GPT, LlaMa 3, & Co. can be prompted to follow specific task instructions, including translation between languages (cf. Lyu *et al.*, 2024). Furthermore, similar to open-source MT models, many LLMs are "open-weights"[10] and thus also facilitate transparent and reproducible research. However, using LLMs for MT compares negatively to using the open-source MT models we consider in terms of resource requirements. Competitive LLMs have many more parameters than the open-source MT models we consider. For example, even the relatively large M2M model has only 1.2 billion parameters compared to the 32 to 70 billion parameters of medium-sized multilingual LLMs. This means that processing and generating a token with an LLM expends much more energy—and thus emits more CO_2—than using a specialized MT model. The size of LLMs, moreover, likely makes local deployment infeasible for most applied researchers, which means that they have to fall back to using a commercial provider's API or hardware.

3. Two studies on the comparative reliability of open-source machine translation

We believe the benefits of using open-source MT models for political science research outweigh the potential costs. However, researchers' main concern should be whether they enable reliable cross-lingual quantitative text analysis in comparison to commercial services (cf. Vries *et al.*, 2018; Reber, 2019; Windsor *et al.*, 2019; Courtney *et al.*, 2020).

We address this question empirically by comparing results from open-source MT against commercial MT. The general intuition of our empirical strategy is simple. We first apply different text-as-data methods to texts that were translated with commercial MT services, open-source MT models, and, if available, by expert translators. We then compare the methods' outputs obtained for the same documents using different translations. This allows us to assess whether translating with an open-source model instead of a commercial service yields systematically different measurements and, hence, levels of reliability.

We cover two widely used quantitative text analysis methods: topic modeling and supervised text classification. Our analyses are based on corpora from different domains of political communication (parliamentary speech, party manifestos, and social media). This broad scope makes our results relevant to a wide range of research applications.

3.1. Study 1: cross-lingual topic modeling

In our first study, we build on de Vries *et al.* (2018) to assess whether the *translation source* (commercial vs. free MT) affects the reliability of LDA topic modeling. The main strength of de Vries and colleagues' original research design is that it allows one to compare the inputs and outputs of a

[9] The application and code are available online.

[10] This means that their parameters are freely available for download and use in (scientific) work, but not the source code or data used for training them.

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

topic model obtained from machine-translated texts to those obtained from human experts' translations. This strategy provides an ideal comparison if we consider human experts "gold standard" translators. For both corpora, de Vries, Schoonvelde, and Schumacher pre-processed the text data,[11] created term-document matrices (TDMs) that count the number of occurrences of words (terms) in a speech (document), and fitted an LDA topic model (Blei *et al.*, 2003). The authors then compare whether the topic content of the models fitted to human and Google Translate translations is similar in terms of word stems and whether the same documents have similar probabilities of being in a specific topic. This empirical strategy allows them to assess how (dis)similar the DFM and topic model outputs obtained from machine-translated documents are compared to those obtained from the human expert-translated corpus. We take advantage of the strength of their study in our analysis.

3.1.1. Empirical strategy

Like de Vries *et al.* (2018), we compare the reliability of MT for topic modeling by comparing topic model outputs to those of a model fitted directly to texts translated by human experts. The intuition of this comparison is the following. If we find higher discrepancies in the model fitted to human translations when we use an open-source MT model for MTs instead of Google Translate, this would indicate that relying on open-source models for translation impairs the reliability of LDA topic modeling. This finding would support the conclusion that open-source MT models' translation quality is insufficient for applied bag-of-words text analysis. However, if the topic model fitted on open-source MT models' translations fares no worse (or even better) than the benchmark model compared to an equivalent topic model fitted on translations obtained with Google Translate, we would conclude that open-source MT models enable comparatively reliable topic modeling of machine-translated texts.

Accordingly, we evaluate whether using open-source MT models for translation instead of Google Translate negatively affects topic models' measurements compared to the human translation benchmark. To compare the quality of open-source MT, we translated the parallel corpora from the respective language into English with OPUS-MT (Tiedemann and Thottingal, 2020).[12] We then pre-processed the data in the same way as de Vries *et al.* (2018) did, created a TDM, and fitted an LDA topic model with 90 topics (using the same random seed, burn-in time, and number of iterations). Finally, we compare the inputs and outputs of topic models fitted to OPUS-MT or Google Translate translations to the ones obtained from human expert-translated texts.[13] We used the same metrics as de Vries *et al.* (2018): The similarity between documents' bag-of-words representations. The similarity between the documents' estimated topic proportions. And the similarity between the estimated topics' document compositions. Overall, our empirical strategy allows us to evaluate whether, compared to the commercial Google Translate translations, the translations obtained from OPUS-MT led to stronger discrepancies vis-à-vis the expert translation benchmark.

To compare Google Translate and OPUS-MT, we compute how much the measurements obtained with the LDA topic models fitted to their translations differ from those obtained with the LDA topic model fitted to expert translations. We then assess how much these deviations from the "gold standard" model differ between the commercial and open-source MT-based topic models. However, statistical significance in difference-in-means tests does not answer the question of whether this particular difference is substantial.[14] We thus also rely on equivalence testing (Lakens *et al.*, 2018). In equivalence testing, researchers define the smallest effect size of interest before analyzing the data.

[11] Deleting speeches containing less than 50 words, removing stop words, numbers, and punctuation, as well as stemming and lower casing words.

[12] We chose OPUS-MT because it had the lowest translation costs regarding time and CO_2 emissions produced by the GPU compute time. This is because OPUS-MT is a suite of language pair-specific models and thus has far fewer parameters than M2M.

[13] To make topic models' outputs comparable, we have matched topics based on the stems allocated to the topics following the procedure adopted by de Vries *et al.* (2018).

[14] Large sample sizes can easily detect statistically significant but very small differences, making it unclear whether such differences should be given much consideration.

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

Table 2. Summary statistics of cosine similarities between bag-of-words representations obtained from machine- and human-translated texts at the document level. Columns grouped by translation model

		Google Translate				OPUS-MT			
Language	*N*	Mean	Std. dev.	Min	Max	Mean	Std. dev.	Min	Max
Danish	2301	0.915	0.063	0.549	0.992	0.902	0.079	0.434	0.988
German	2148	0.915	0.074	0.488	0.991	0.920	0.072	0.279	0.992
Spanish	2335	0.929	0.059	0.483	0.991	0.935	0.056	0.614	0.992
French	2347	0.925	0.064	0.564	0.989	0.930	0.060	0.510	0.991
Polish	2338	0.913	0.073	0.475	0.989	0.908	0.085	0.130	0.990
Total	11469	0.920	0.067	0.475	0.992	0.919	0.072	0.130	0.992

The observed difference in means is then compared with this threshold to examine whether an effect size falls within a range of values considered equivalent to a null effect. If the test statistic and its confidence interval exceed the range of values defined as negligible, the equivalence hypothesis is rejected. We set the equivalence bound to ± 0.01.

3.1.2. Results

Overall, our results indicate that using the open-source OPUS-MT model for translation instead of Google Translate does *not* negatively affect topic models' reliability compared to the human translation benchmark.

First, we look at how the MT source affects the bag-of-words inputs to the topic model. Specifically, we compare the TDM obtained from the human expert translations with the TDMs obtained from Google Translate and OPUS-MT at the document level by computing document-level cosine similarity scores. As summarized in Table 2, Google Translate and OPUS-MT result in very similar document input representations (see also Figure B1 in the Supplementary Material). Specifically, the two MT models are on par when comparing the total means of TDMs cosine similarity with the human translation benchmark. As with Google Translate, 92 percent of the OPUS-translated documents had a cosine similarity of 0.8 or higher with the human-translated English texts. Therefore, using OPUS-MT for MT resulted in document representations that are as similar to those obtained from human expert translations as when using Google Translate.

The discrepancies between the bag-of-words obtained by tokenizing texts translated with OPUS-MT instead of Google Translate seem to be explained by these models' different vocabulary sizes. Upon closer inspection of the TDMs, it appears that both the texts translated by human experts and those translated with Google Translate contain a similar number of unique terms. In contrast, the number of unique terms in the corpus translated with OPUS-MT is lower (see Table B1 in the Supplementary Material). As a result, the overlap of unique terms between the human gold standard is slightly smaller for OPUS-MT than for Google Translate per language. This might be because OPUS-MT is an open-source model, and its English vocabulary is thus likely more limited than that of Google Translate's model. This limitation in vocabulary size may explain other differences between Google Translate and OPUS-MT, as overlap in word frequencies is also important for topic modeling tasks.

As the next step of our analysis, we assess how the MT source affects documents' estimated topic proportions, one of the main outputs of the LDA topic model. As shown in Table 3, we find that documents' estimated topic proportions are highly similar between Google Translate and the gold standard model and between the gold standard model and OPUS-MT (see also Figure B2 in the Supplementary Material). The average correlation between the measurements of the topic model using OPUS-MT translations and those of the model fitted to human translations is 0.785. For the model based on Google Translate translations, this correlation is 0.782. Looking at differences across languages, it is notable that OPUS-MT performs well for French, with an average correlation of 0.801. On the other hand, Google Translate has a slightly higher correlation with the gold standard for German

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

Table 3. Summary statistics of correlations between document-level topic proportion estimates obtained from machine- and human-translated texts. Columns grouped by translation model

		Google Translate				OPUS-MT			
Language	*N*	Mean	Std. dev.	Min	Max	Mean	Std. dev.	Min	Max
Danish	2301	0.809	0.237	−0.059	0.999	0.788	0.185	−0.075	0.996
German	2148	0.799	0.156	0.007	0.997	0.780	0.187	0.035	0.997
Spanish	2335	0.772	0.211	−0.092	0.997	0.781	0.191	−0.048	0.997
French	2347	0.761	0.194	−0.069	0.996	0.801	0.206	−0.036	0.997
Polish	2338	0.769	0.218	−0.038	0.995	0.778	0.243	−0.052	0.997

and Danish, with an average correlation of 0.799 and 0.809, respectively. Overall, however, as these differences between the two models are small, we conclude that there are *no* substantial differences in topic proportions assigned to documents when using OPUS-MT instead of Google Translate.

Lastly, we assess whether the topics learned by the models trained on machine-translated texts are comparable to those learned by the model using human expert translations. As shown in Figure 1, the correlations of learned topics' prevalence across documents with those in the human translation benchmark are overall very high for both MT models. This holds regardless of the source language, and we hence did not observe any language bias when comparing the performance of OPUS-MT and Google Translate.[15] And as shown in Figure B3 in the Supplementary Material, there is also no substantial difference between MT models in terms of how the topics learned from their translation compared to the human translation benchmark in terms of content. The stems load similarly on topics extracted from the machine-translated texts as on the same topics extracted from human-translated data for French, Danish, and German. However, slight differences exist in the overlap of features for Polish and Spanish. Overall, the results do not indicate a substantial difference between Google Translate and OPUS-MT in their alignment with the word stems of topics in the human translation-based topic models, except for a minor difference for Danish speeches (see Table B2 in the SM).

3.2. Study 2: machine translation in supervised classification

Our second study examines the role of MT in transformer-based supervised text classification. We assess whether researchers should expect supervised classifiers to be less reliable in a labeling task when fine-tuned using texts translated with an open-source MT model instead of a commercial one.

We address this question with a large comparative classifier fine-tuning experiment. Our experiment allows us to estimate the expected differences in predictions for held-out texts and overall reliability between classifiers fine-tuned using an open-source MT model's translations as text inputs to their counterparts fine-tuned with translations generated by a commercial MT model.

3.2.1. Data

We have compiled a large benchmark of human-labeled multilingual political text datasets for our analyses. Our benchmark includes five datasets (Theocharis *et al.*, 2016; Lehmann and Zobel, 2018; Düpont and Rachuj, 2022; Poljak, 2023; Ivanusch and Regel, 2024). As shown in Table 4, these datasets jointly cover 31 languages from the Indo-European, Uralic, and Turkic language families and three domains of political communication (party manifestos, parliamentary speech, and social media).

We have obtained MTs into English for all non-English texts in our datasets. The CMP Translations corpus contributed by Ivanusch and Regel (2024) contains English translations of sentences generated with DeepL, and we have added translations from two open-source MT models (M2M 1.2B and

[15]Both translation systems excel in Danish and Romanic languages and perform slightly less well on Polish, the one Slavic language in our parallel corpus.

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

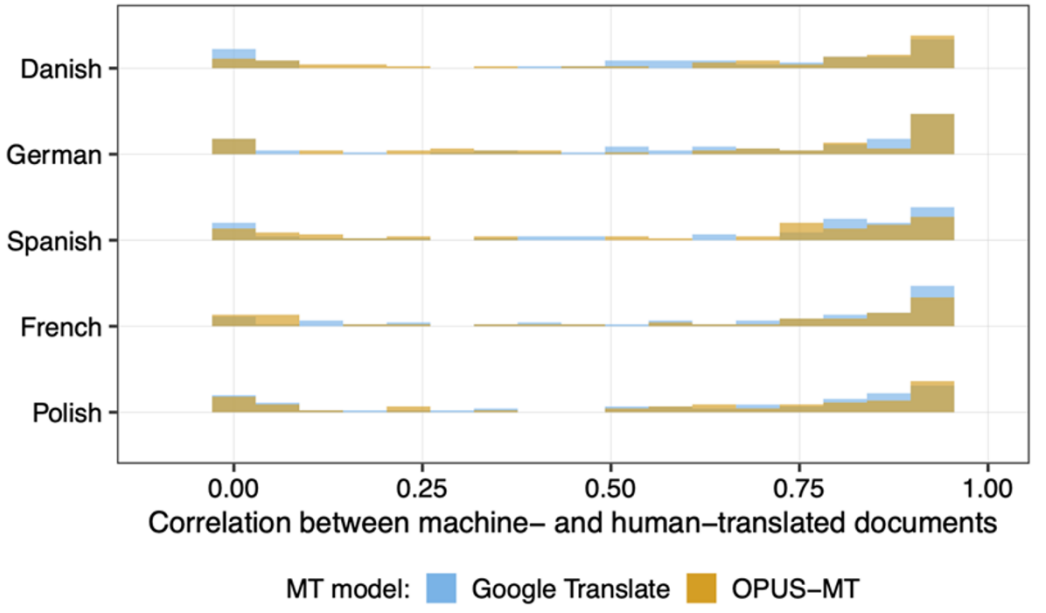


Figure 1. Similarity of corpus-level topical prevalence between the human translated corpus and the Google Translate and OPUS-MT, respectively. Topical prevalence means the probability of the document being assigned to a specific topic.

Table 4. Datasets

Description	Languages
Ivanusch and Regel (2024): language-stratified sample of sentences taken from the CMP Translations corpus	Bulgarian, Czech, Danish, Dutch, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Lithuanian, Norwegian, Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish, Turkish, Ukrainian
Düpont and Rachuj (2022): sentences from manifestos taken from the CMP corpus	Danish, Dutch, Finnish, French, German, Italian, Spanish, Swedish
Lehmann and Zobel (2018): quasi-sentences from manifestos taken from the CMP corpus	Danish, Dutch, English, Finnish, French, German, Spanish, Swedish
Poljak (2023): parliamentary speeches delivered in Question Time sessions	Dutch, English, French [a]
Theocharis *et al.* (2016): tweets, retweets, and replies to tweets by candidates for the 2014 European Parliament election	English, German, Spanish [b]

[a]Additional analyses conducted including speeches from Croatia written in Bosnian (bs) and Croatian (hr).
[b]Additional analyses conducted including tweets written in Greek (el).

OPUS-MT).[16] For all other datasets, we have obtained English translations from two commercial services (DeepL and Google Translate) and three open-source MT models (M2M 418 M, M2M 1.2B, and OPUS-MT). In addition, all but the Lehmann and Zobel (2018) data come with Google Translate translations that the datasets' owners obtained when compiling them.[17]

The texts in the five datasets included in our benchmark have been coded on multiple dimensions. As shown in Table 5, our benchmark thus covers various target concepts, ranging from policy topics to negative campaigning to incivility. We use this variation within and across datasets to define 15

[16]We have subset the CMP Translations corpus to full sentences (using regular expressions) and sampled 2707 sentences per language to balance the language distribution.

[17]Theocharis *et al.* (2016) obtained their Google Translate translations in 2015, Düpont and Rachuj (2022) in 2020, and Poljak (2023) in 2021, according to our correspondence with the authors.

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

Table 5. Tasks overview

Ivanusch & Regel: sample of sentences taken from the CMP Translations corpus	
Classify the policy topic discussed in quasi-sentence	Extrel, freedem, polsys, econ, welqual, fabsoc, socgrp
Classify the stance expressed in quasi-sentence	Left, right, none
Classify the stance expressed in quasi-sentence (binary)	Left, right
Classify the stance expressed in quasi-sentences about economic issues (binary)	Left, right
Düpont & Rachuj: sentences from manifestos taken from the CMP corpus	
Classify the policy topic discussed in quasi-sentence	Extrel, freedem, polsys, econ, welqual, fabsoc, socgrp
Classify the stance expressed in quasi-sentence	Left, right, none
Classify the stance expressed in quasi-sentence (binary)	Left, right
Classify the stance expressed in quasi-sentences about economic issues (binary)	Left, right
Classify the stance expressed in quasi-sentences about the issue of freedom and democracy (binary)	Left, right
Lehmann & Zobel: quasi-sentences from manifestos taken from the CMP corpus	
Classify the issue focus of quasi-sentences	Immigration, integration
Classify the stance expressed in quasi-sentences about the issues of immigration and integration (binary)	Position: sceptical, supportive
Poljak: parliamentary speeches delivered in Question Time sessions	
Detect whether a speech contains one or more attacks by parliamentary actors	Attack: yes, no
Classify the type of attack	Attack type: policy, trait, both
Detect uncivil attacks	Uncivil: yes, no
Theocharis et al.: tweets, retweets, and replies to tweets by candidates for the 2014 European Parliament election	
Classify the sentiment of tweets	Positive, neutral, negative
Classify the sentiment of tweets (binary)	Positive, negative
Classify the type of communication in tweets	Broadcasting, engaging
Detect impolite tweets	Polite, impolite
Detect political tweets	Political: yes, no

classification tasks (see column two in Table 5). We adopt this comparative approach to facilitate the generalizability of our findings. By including tasks focusing on concepts with varying levels of difficulty in corpora from different domains, we ensure that our findings on the "reliability cost" of using open-source MT are not specific to a single dataset, target concept, or classification task.

3.2.2. Empirical strategy

We proceed in two steps. First, we analyze the similarities of translations obtained with open-source and commercial MT models. This enables us to understand how using an open-source MT model instead of a commercial one might impact the inputs used to fine-tune a classifier.

Second, we turn to our main research question of whether using an open-source MT model instead of a commercial one to translate the texts taken as inputs when fine-tuning impacts a classifier's predictions and reliability negatively. For this analysis, we fine-tune one text classifier per MT model, dataset, and task covered by our benchmark (see Table 5) on a training data split. This results in four classifiers per task in the CMP Translations corpus and six (five) classifiers per task for the other datasets with(out) old Google Translate translations.[18] In addition, we have fine-tuned one

[18]We have used the RoBERTa base checkpoint for fine-tuning for all except those in the Theocharis *et al.* (2016) dataset. As the Theocharis *et al.* (2016) dataset records Twitter posts, we used the RoBERTa-based Twitter language model pretrained by Barbieri *et al.* (2020) instead.

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

multilingual classifier per task to enable comparisons with the alternative strategy of aligning texts by embedding instead of translation (cf. Licht and Lind, 2023).[19]

We then use these classifiers to predict the labels of held-out texts in the corresponding test set splits. Holding constant the random seed, data splits, and training hyper-parameters used when fine-tuning,[20] the only factor that varies between classifiers fine-tuned for the same task is the translation model used to translate the source texts into English. We can thus attribute differences in classifiers' predictions and out-of-sample classification performances to the MTs used to fine-tune them.

Specifically, we compare classifiers at two levels of analysis. First, we compare classifiers' language- and label-class-specific reliability in labeling held-out texts using the F1 score.[21] Second, we analyze the agreement and discrepancies at the level of predicted labels of the test set examples.

3.2.3. Similarities between open-source and machine commercial translations

A first step to understanding how using an open-source MT model instead of a commercial one might impact the predictions and classification reliability of a classifier fine-tuned using these translations as inputs is to assess how similar or dissimilar the translations generated with alternative models are.

We address this question based on a sample of 500 sentences per language in our full-sentence sample of the CMP Translations corpus[22] by computing the similarity between translations obtained with M2M (1.2B), respectively, OPUS-MT, to the translations obtained with DeepL for each sentence in this sample.[23] To measure the similarity between two translations of the same source sentence obtained with different MT models, we rely on BERTScore (Zhang *et al.*, 2020). BERTScore is a method for estimating the semantic similarity of two sentences by comparing the contextualized embeddings of their tokens obtained with a pre-trained transformer model.[24]

Figure 2 summarizes the distribution of BERTScore F1 scores for translations obtained with the M2M (1.2B) and OPUS-MT models compared to DeepL translations by language (y-axis) and open-source MT model (plot panels). Numbers plotted above box plots report average BERTScore F1 scores per language and open-source MT model. The closer the BERTScore F1 is to 1.0, the more similar a translation pair is.

Figure 2 shows that both open-source MT models typically generate translations highly similar to DeepL translations, as translations' similarities are mostly between 0.9 and 1.0. This indicates that the two open-source MT models typically generate translations that are semantically highly similar to those produced by DeepL.

However, three observations are notable. First, as further illustrated in Figure C1 in the Supplementary Material, for each of the two open-source MT models we analyze, there are some languages whose translations' similarities to DeepL translations are, on average, above the cross-language average (i.e., grand mean) of all languages, and some languages whose average translation similarities are below this cross-language average. For example, the similarity of the M2M (1.2B) model's translations of Spanish and Swedish sentences to their DeepL translations is above average, while its Finnish, Hungarian, and Turkish translations are below average. Second, this sorting of languages into above- and below-average groups is partially consistent across open-source MT models (e.g., Hungarian, Spanish, Swedish, and Turkish). Assuming that DeepL yields on average higher quality translations,

[19] Using XLM-T (Barbieri *et al.*, 2022) for the tasks in the Theocharis *et al.* (2016) dataset and XLM-RoBERTa (Conneau *et al.*, 2020) instead.

[20] Note that we applied inverse class weighting to improve classification performance for minority label classes, and for severely imbalanced tasks, we down-sampled majority class instances in the training data.

[21] We focus on the F1 score because the label classes are imbalanced in all our tasks (cf. Table C4–C7).

[22] We use the CMP Translations corpus in this analysis because it offers the largest language coverage among our labeled datasets in our benchmark.

[23] We limit our focus to these models because our analysis in Section 3.2.4 suggests that DeepL is a suitable benchmark for high-quality translation, and the smaller M2M (418 M) model underperforms.

[24] Alternative metrics, like BLEU or the Jaccard similarity, measure the overlap of n-grams between sentences and thus only indicate lexical similarity.

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

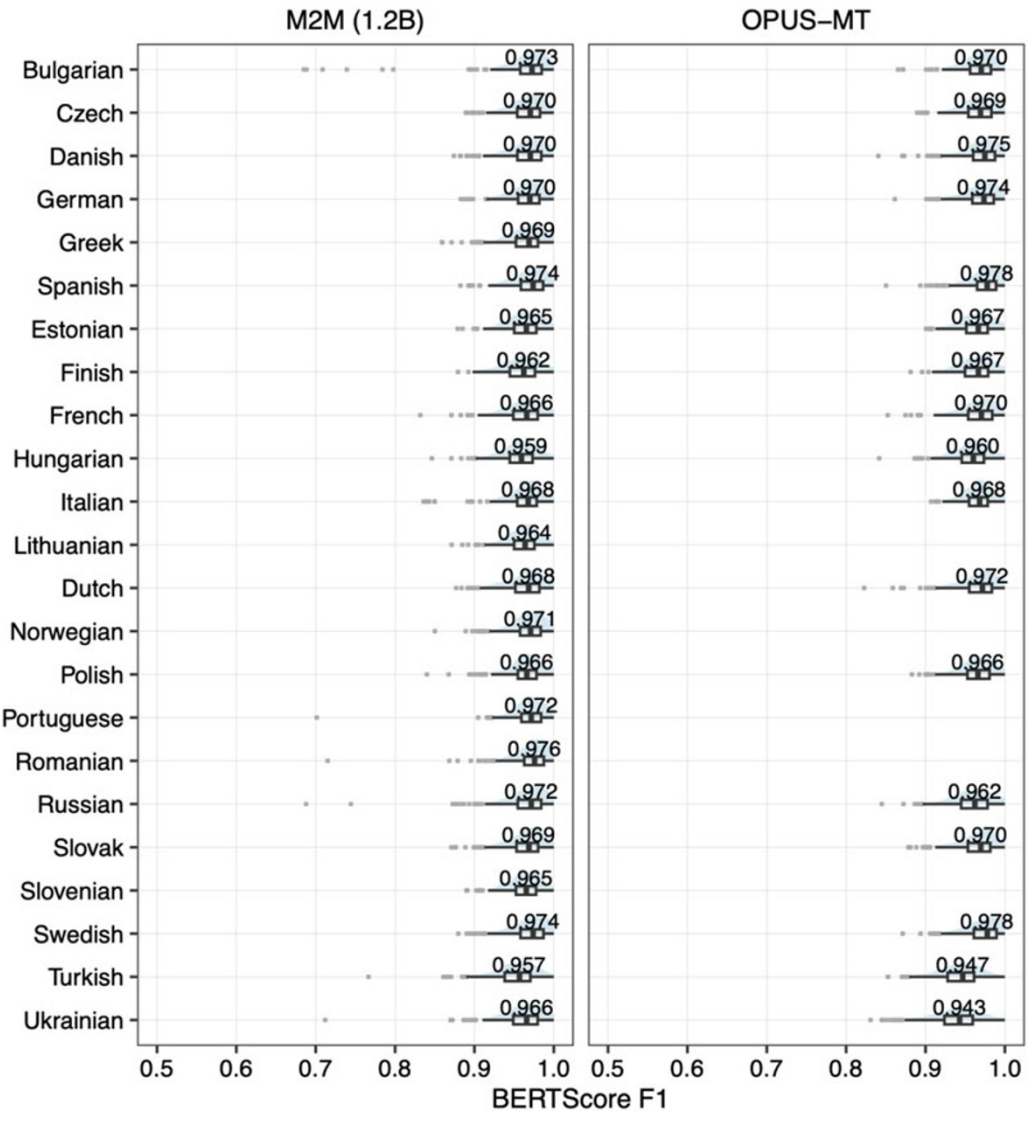


Figure 2. Distribution of similarities of open-source MT models' translations to DeepL translations by language and open-source MT model in sample of 500 sentences per language sampled from the CMP Translations corpus. Translation similarity measured with BERTScore at translation pair level. Note that no OPUS-MT translation to English was obtained for Greek, Lithuanian, Norwegian, Portuguese, Romanian, and Slovenian due to translation direction limitations.

this suggests that the two open-source MT models share some strengths and weaknesses. Third, while the estimated pairwise similarity is above 0.9 for the vast share of sentences, there are a few "outlier" data points with BERTScore F1 scores below the lower whisker of the box plot for each language and open-source MT model. These are cases of high semantic discrepancy between open-source MT and DeepL translations, and we will pay greater attention to these outliers in an analysis reported in Section 3.4.

3.2.4 (No) reliability cost of open-source machine translation

Next, we compare the reliability of classifiers fine-tuned using open-source MT models' translations to classifiers fine-tuned using commercial MT models' translations. Specifically, we compare

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

language- and label-class-specific F1 scores between classifiers trained on translations generated with different MT models within dataset, task, and language through regression modeling and complement these analyses with equivalence testing. While there is much heterogeneity in classifiers' performance across datasets, tasks, and languages, our experimental setup allows us to focus on the translation model (type) as a factor shaping classifiers' behavior. However, a detailed breakdown of classifiers' F1 scores by dataset, task, label class, and/or language can be found in Table C9, Figures C2–C6, and Tables C10–C14 in the Supplementary Material.

We first present the results of the regressions that model the overall effect of the (type of) MT model used to translate texts on classifiers' language- and label-class-specific F1 scores, and add interactions with source language indicators in a second step to assess heterogeneous effects across language.[25] To account for uncertainty in estimates of classifiers' test set F1 scores, we bootstrap F1 estimates from each classifier's test set predictions and use these bootstrapped scores in our analysis. Furthermore, to account for heterogeneity in F1 scores across datasets, tasks, languages, and label classes (cf. Figures C2–C6), all our regressions include dataset, task, source language, and label class fixed effects and cluster standard errors by source language, tasks, label class, and translation model.

Specifically, we estimate the following regressions:

$$\mathrm{F1}_d, t, c, l = {}^{\beta}0 + {}^{\beta}1\text{MT model type} + {}^{\delta}d + {}^{\theta}t + {}^{\kappa}c + {}^{\lambda}l + {}^{d}, t, c, l \tag{1}$$

$$\mathrm{F1}_d, t, c, l = {}^{\beta}0 + {}^{\beta}1\text{MTmodel} + {}^{\delta}d + {}^{\theta}t + {}^{\kappa}c + {}^{\lambda}l + {}^{d}, t, c, l \tag{2}$$

$$\mathrm{F1}_d, t, c, l = {}^{\beta}0 + {}^{\beta}1\text{MT model} + {}^{\beta}2\text{language} + \beta 3(MTmodel \times language) + \delta d + \theta t + \kappa c + \lambda l + \epsilon\, d, t, c, l \tag{3}$$

where F1 score $_{d,\,t,\,c,\,l}$ represents a bootstrapped F1 score for dataset d, task t, label class c, and language l; in equation (1), "MT model type" is a categorical indicator differentiating between commercial MT-based classifiers (reference category), open-source MT-based translations, and multilingual Transformer-based classifiers; in equations (2) and (3), "MT model" indicates the MT model used for translation (using DeepL as the reference category); in equation (3), "language" indicates the source language; and δ, θ, κ, and λ are the dataset, task, label class, and source language fixed effects estimates, respectively.[26]

If for model (1) our estimate of β_1 for the open-source MT-based classifier category is negative, this would indicate that using an open-source MT model for translation instead of a commercial one results, on average, in a lower F1 score, indicating poorer reliability in labeling held-out texts. The magnitude of the coefficient estimate, in turn, indicates how much worse.

We first turn to the overall effect of open-source versus commercial MT (equation 1). Model 1 in Table 6 estimates the average difference in classifiers' bootstrapped language and label-class-specific F1 scores when using an open-source MT model instead of a commercial one. This difference is estimated to be negative and statistically significant ($t = -9.57$, $p < 0.000$). However, the estimated magnitude is only 0.009, that is, less than a difference of 0.01 units on the [0, 1] F1 score scale.

[25] For datasets that contain English texts, we exclude F1 scores in classifying English-language texts from this comparison.

[26] We restrict our analyses to the languages that can be translated into English by all models, as our goal is to compare the classifiers fine-tuned for each task using the translations obtained with different MT models. For example, we excluded the Greek tweets in the Theocharis *et al.* (2016) dataset because they cannot be translated to English by the OPUS-MT model (see the note in Table 4). Moreover, for model (3), we subset the data to classifiers fine-tuned using input texts' translations from DeepL, Google Translate, M2M (1.2B), or OPUS-MT.

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

Table 6. OLS coefficient estimates of the effect of using open-source vs. commercial machine translation models for translating input texts on classifiers' language-specific out-of-sample classification performance (F1 score)

	Model 1	Model 2
Intercept	0.578 (0.003)***	0.579 (0.003)***
Type of MT model (ref.: commercial MT model)		
open-source MT model	−0.009 (0.001)***	
Translation model (ref.: DeepL)		
Google Translate		0.002 (0.002)
Google Translate (old)		−0.003 (0.002)
OPUS-MT		−0.002 (0.002)
M2M (1.2B)		−0.006 (0.001)***
M2M (418 M)		−0.014 (0.001)***
Multilingual classifier	−0.020 (0.001)***	−0.020 (0.001)***
R^2	0.349	0.350
Adj. R^2	0.349	0.350
Num.obs.	103500	103500

***$p < 0.001$; ** $p < 0.01$; *$p < 0.05$.
The F1 score is measured on a scale from 0 to 1. A coefficient estimate of, for example, +0.01 (+0.001) represents an average increase of the F1 score by 0.01 (0.001), that is, one (a tenth of one) F1 score point. All models include data set, task/outcome, and language fixed effects. Standard errors clustered by data set, task/outcome, language, and, in case of tasks with more than two labels, by label class.

Thus, when fine-tuning a supervised text classifier, researchers should expect a reduction in its out-of-sample classification reliability if they use an open-source instead of a commercial MT model for translation. However, they can expect that this reduction will be negligibly small, considering that even classifiers fine-tuned on different folds of the same dataset (Licht, 2023; Laurer *et al.*, 2024) or with different seeds (Wang, 2023) usually exhibit higher levels of variability in test set F1 scores than our estimate of 0.009.

Equivalence tests (Lakens *et al.*, 2018) support this conclusion (see Table C17 in the Supplementary Material). The difference between the average bootstrapped F1 scores of commercial and open-source MT-based classifiers is 0.006 and within the equivalence bound of [−0.02, 0.02] ($t = -5.338$, $p < 0.000$). Moreover, Model 1 in Table C15 in the Supplementary Material shows that these findings hold when we omit classifiers fine-tuned with input text translations generated with older versions of Google Translate from the comparison. And Model 2 in Table C15 presents evidence that the estimated F1 score difference of 0.009 reported in Model 1 in Table 6 drops by 53% if we remove classifiers fine-tuned with input text translations generated with the small (418B) M2M model from the comparison.

Model 2 in Table 6 adds nuance to these findings. It compares classifiers fine-tuned with DeepL translations to ones fine-tuned using one of the other MT model's translations. This shows that using the open-source OPUS-MT instead of DeepL does not significantly reduce classifiers' test set F1 scores. In the case of the M2M (1.2B) model, the difference is negative and statistically significant, but with a magnitude of only 0.006 on the [0, 1] F1 score scale, again small and arguably practically negligible.[27] Furthermore, Model 2 in Table 6 shows that even the "worst" alternative of using the smaller M2M (418 M) model only reduces classifiers' test set F1 score by 0.014 compared to using DeepL.

Turning to cross-language heterogeneity, Figure 3 reports predicted values of the regression model fitted with equation (3) on page 19.[28] It shows that our finding of a very small reliability cost of open-source MT in supervised text classification holds for most high- and low-resource languages included

[27]Equivalence tests for the MT model-specific comparisons show that equivalence holds in the ± 0.02-bound for all comparisons except those between Google Translate and the M2M models and the comparison between DeepL and M2M (418 M). In the case of comparisons with DeepL-based classifiers, we even find support for equivalence in the ± 0.01-bounds for OPUS-MT-based classifiers ($t = -3.447$, $p < 0.000$).

[28]Table C18 reports the regression coefficient estimates.

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

in our benchmark.[29] In high-resource languages (top panel), open-source MT-based classifiers are predicted to perform, on average, slightly worse for Danish, German, Spanish, Dutch, and Swedish and better for French and Italian. However, these differences are within ± 0.01 on the [0, 1] F1 score scale for all these languages except Swedish. For low-resource languages (bottom panel), this pattern replicates in many languages, including those with other scripts (e.g., Bulgarian, Greek, and Russian) and/or from other language families than the Romance and Germanic families (Czech, Estonian, Finnish, Hungarian, Romanian, and Slovenian). Comparatively large negative differences (≤ -0.015) exist only for Norwegian, Polish, Slovak, Turkish, and Ukrainian. However, it is worth emphasizing that some of these languages are only represented in one dataset of our benchmark (see Table 4), which limits our ability to generalize the language-specific patterns we find for these languages.

3.3. Example-level prediction agreement and win rates

The previous analysis has an important limitation. Classifiers can achieve a similar F1 score in a task even if they yield different predictions for individual examples. The F1 scores as a measure of predictive reliability might thus mask more nuanced example-level differences between classifiers.

To address this concern, we refocus the above analysis by using a measure of classifiers' prediction-level pairwise agreement instead of the F1 score to compare MT models and their effect on classifiers' classification behavior. Specifically, we compare the test set examples' predicted labels between two classifiers trained using translations obtained with different MT models using DeepL-based classifiers as a reference.[30] This analysis shows that the labels predicted by OPUS-MT and M2M (1.2B)based classifiers agree to a similar extent with their DeepL-based counterparts' predicted labels as Google Translate-based classifiers (see Table C20 and Figure C7 in the Supplementary Material). Specifically, the average agreement of Google Translate-based classifiers predicted labels with those predicted by their DeepL-based counterparts is 0.836. The agreement rates of the OPUS-MT and M2M (1.2B)-based classifiers are estimated to be only −0.011 lower, and this difference is found to be statistically insignificant.[31]

3.4. Classification accuracy and translation quality

In a supplementary analysis reported in detail in Section C.4.2 of the Supplementary Material, we further analyze whether low-quality translation might be associated with less reliable classifications. This analysis is motivated by Figure 2, which shows that open-source and commercial models' translations can diverge substantially in rare cases and therefore raises the question of whether we observe higher classification error rates in texts with strongly divergent translations compared to texts that the MT models translate very similarly.

We find that for both DeepL and M2M (1.2B)-based classifiers, unilateral classification errors are more likely in sentences whose translations diverge comparatively strongly between DeepL and M2M (1.2B). Importantly, this tendency is slightly more pronounced for classifiers relying on M2M's open-source translations. If we assume that in cases where translations diverge, DeepL yields the higher-quality translation, this finding suggests that using open-source MT models for translations can indeed negatively affect the classification quality when the underlying MT model generates poor translations.

[29] We have obtained predicted F1s for all combinations of MT model type (commercial vs. open-source), languages, datasets, and tasks, and averaged the predicted values and 95% confidence intervals by MT model type and language.

[30] This approach has the additional advantage that we do not need to treat the human-generated test set labels as "ground truth," although we know that human annotations can be quite noisy.

[31] Another way to learn about the impact of using open-source instead of commercial MT models to translate the inputs used to fine-tune text classifiers is to study cases of disagreements between classifiers' predicted labels. We report this analysis in Section C.4.1 of the Supplementary Material.

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

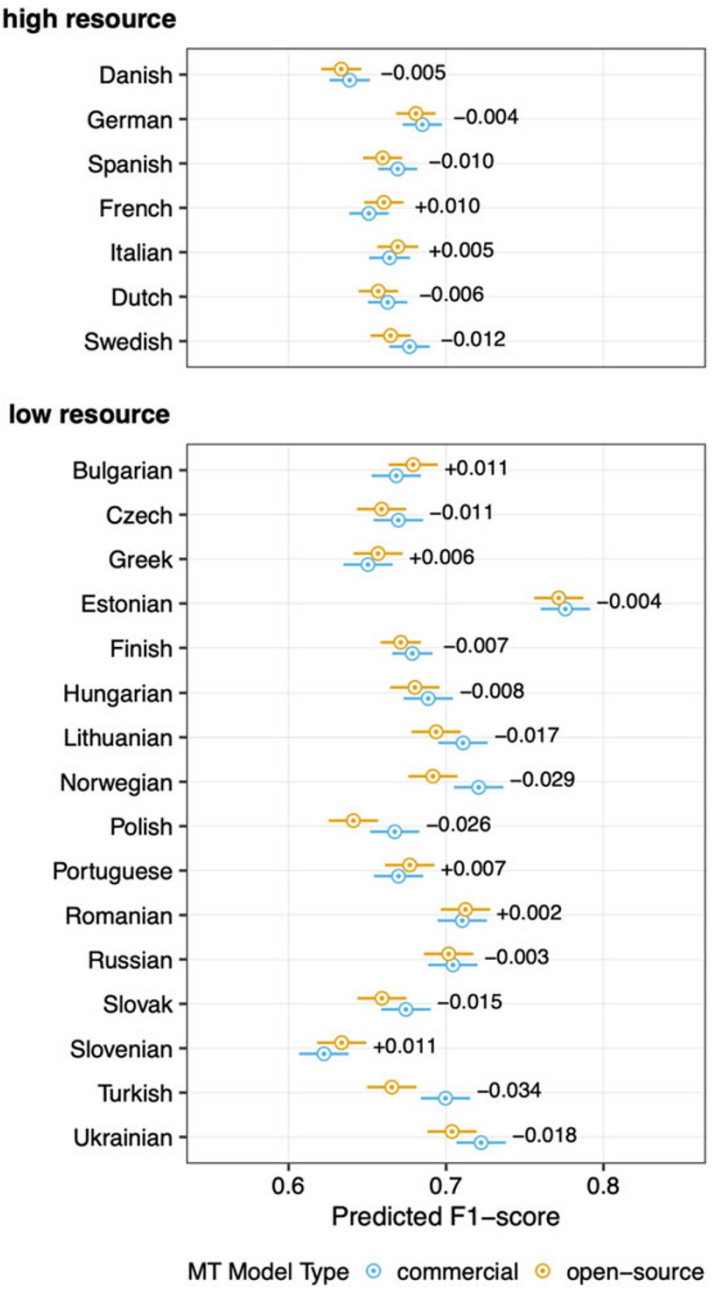


Figure 3. Predicted language-specific F1 scores by language and type of MT model. Estimates based on regression reported in Table C18.

However, our analyses also show that the differences between DeepL and M2M-based classifiers' accuracy at low levels of translation similarity and the differences between accuracies at low and high levels of translation similarity are very small. Moreover, we have shown in Figure 2 that cases of high disagreement between commercial and open-source MT models' translations are rather rare. Thus, even if open-source MT models might occasionally yield comparatively poor-quality translations,

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

this should impair supervised classifiers' reliability only in the margins. Yet, we acknowledge that our evidence on the question of whether low-quality open-source translations impair classifiers' measurement reliability more than in the case of using a commercial MT model would require further support to justify a conclusive judgment, as our analysis lacks direct indicators of translation quality or error.

4. Conclusion and discussion

Open-source MT models such as OPUS-MT (Tiedemann and Thottingal, 2020) and M2M (Fan *et al.*, 2021) are transparent, reproducible, and affordable alternatives to commercial MT services like Google Translate and DeepL. We have assessed whether translating multilingual corpora into English with available open-source MT models instead of a commercial service (Google Translate) yields substantially different results when applying two common computational text analysis methods. Our first study extends the study by Vries *et al.* (2018), who evaluated MT for cross-lingual topic modeling for four Indo-European languages. Our second study is an original analysis of the reliability of open-source MT for cross-lingual supervised text classification with transformer-based classifiers based on a large benchmark of human-annotated texts covering Germanic, Romance, Slavic, Indo-European, Uralic, and Turkic languages.

Our findings support the conclusion that open-source MT models can replace commercial services when applying bag-of-words topic modeling and transformer-based supervised text classification. We find only minor differences between the measurements obtained from corpora translated with open-source models and commercial services. In the case of our topic model analyses, we find that the topics estimated by a model fitted on parliamentary speeches we machine-translated with the open-source M2M model are as similar to the benchmark topic model fitted on human-translated speeches as those estimated by its counterpart model fitted on speech translations generated with Google Translate. In addition, both MT-based models allocate speeches to topics similar to those in the benchmark model. In the case of our supervised text classification study, we find that the difference between transformer-based classifiers fine-tuned using translations from open-source MT models performs, on average, only 0.009 F1 scores worse than comparable classifiers fine-tuned using translations from a commercial MT model as input.

Our findings have important implications for applied researchers. Given that using open-source MT models for topic modeling or fine-tuning transformer-based classifiers tends to result in no less reliable measurements than using a commercial MT service, applied researchers can benefit from the transparency and reproducibility advantage of open-source MT. Maybe as important from a practical point of view, using open-source MT models can save researchers costs. For example, relying on an open source instead of a commercial MT service to translate the non-English texts in the four benchmark data sets used in Study 2 would save them about US $ 1267 (see Table C8).

However, our study is not without limitations. First, our analyses focus mainly on MT from Germanic and Romance Indo-European languages. We address this limitation by including the Manifesto Translations corpus (Ivanusch and Regel, 2024) in our supervised classification benchmark, which includes texts written in Slavic Indo-European, Turkic, and Uralic languages. However, future research should extend or complement our benchmark and analyses to Asian and African languages.

Second, as is standard in the methodological literature (cf. Table A1), all our analyses use English as the target language. When researchers study corpora recording only Slavic or Nordic languages, for example, it might be better for measurement reliability to translate texts to the majority language in their corpus or a language from which other languages descended.

Third, researchers might find translation with open-source MT models technically challenging. While software packages such as the EasyNMT Python package provide a handy toolkit,[32] we

[32]See https://easynmt.net/demo.

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

acknowledge that deploying these models and using GPU computing environments are not trivial skills. The code base and online app we provide thus aim to lower this accessibility barrier.[33]

Fourth, researchers might be concerned about the longer processing times for open-source MT models. While we acknowledge that translation with open-source models takes longer than querying commercial services APIs, we note that MT incurs a one-time cost. Many large multilingual, political text corpora, like ParlSpeech2 (Rauh and Schwalbach, 2020) or the Manifesto Project Dataset (Lehmann *et al.*, 2023), are widely used by many researchers. Translating these and other corpora is a one-time investment that would significantly contribute to the research community (e.g., Plenter, 2023).

Fifth, our topic modeling study only examines the LDA topic model, excluding neural topic modeling methods. Similarly, our supervised classification study only examines transformer encoder fine-tuning while ignoring recent developments in using generative LLMs for prompt-based zero- and few-shot political text classification (e.g., Gilardi *et al.*, 2023).

Last but not least, our studies focus on quantitative text analysis. Our research design does not take into account how humans understand machine-translated texts and whether translation errors lead to faulty interpretation, for example, in manual content analysis or text annotation.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2026.10102. To obtain replication material for this article, please visit https://doi.org/10.7910/DVN/8UOOEW.

Acknowledgements. We thank three anonymous reviewers and participants of the EPSA panel ME07 "Multilingual Text Analysis" for helpful comments and suggestions. This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1-390838866.

References

Baden C, Pipal C, Schoonvelde M and van der Velden MACG (2022) Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures* **16**(1), 1–8.

Barberá P, Gohdes AR, Iakhnis E and Zeitzoff T (2024) Distract and divert: How world leaders use social media during contentious politics. *The International Journal of Press/Politics* **29**(1), 47–73 doi:10.1177/19401612221102030.

Barbieri F, Anke LE and Camacho-Collados J (2022) XLM-t: Multilingual language models in twitter for sentiment analysis and beyond. In Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, Goggi S, Isahara H, Maegaard B, Mariani J and Mazo H (eds), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. LREC 2022. European Language Resources Association, pp. 258–266.

Barbieri F, Camacho-Collados J, Anke LE and Neves L (2020) TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Cohn T, He Y and Liu Y (eds), *Findings of the Association for Computational Linguistics: EMNLP 2020*. Findings 2020. Online: Association for Computational Linguistics., pp. 1644–1650.

Baum MA and Zhukov YM (2019) Media ownership and news coverage of international conflict. *Political Communication* **36**(1), 36–63.

Blei DM, Ng AY, Jordan MI and Lafferty J (2003) Latent dirichlet allocation. *Journal of Machine Learning Research* **3**(4), 993–1022.

Chan C-H, Zeng J, Wessler H, Jungblut M, Welbers K, Bajjalieh JW, van Atteveldt W and Althaus SL (2020) Reproducible extraction of cross-lingual topics (rectr). *Communication Methods and Measures* **14**(4), 285–305.

Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L and Stoyanov V. 2020. Unsupervised cross-lingual representation learning at scale. In Jurafsky D, Chai J, Schluter N and Tetreault J (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Association for Computational Linguistics, pp. 8440–8451.

Courtney M, Breen M, McMenamin I and McNulty G (2020) Automatic translation, context, and supervised learning in comparative politics. *Journal of Information Technology & Politics* **17**(3), 208–217.

Dancygier R and Margalit Y (2020) The evolution of the immigration debate: Evidence from a new dataset of party positions over the last half-century. *Comparative Political Studies* **53**(5), 734–774.

Dolinksy A, Schoonvelde M and van der Velden MACG. 2022. Multilingualism in computational text analysis methods: Evidence from a pre-registered survey experiment. In *11th Meeting of the European Political Science Association*. Prague: EPSA 2022.

[33] Available online.

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press

Düpont N and Rachuj M (2022) The ties that bind: Text similarities and conditional diffusion among parties. *British Journal of Political Science* **52**(2), 613–630.
Fan A, Bhosale S, Schwenk H, Ma Z, El-Kishky A, Goyal S, Baines M, Celebi O, Wenzek G and Chaudhary V (2021) Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research* **22**(107), 1–48.
Gilardi F, Alizadeh M and Kubli M (2023) ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* **120**(30), e2305016120.
Ivanusch C and Regel S. 2024. Manifesto corpus translation.
Koehn P. (2005) Europarl: A parallel corpus for statistical machine translation. *Proceedings of Machine Translation Summit X*, Phuket, Thailand, 13–15 September, pp. 79–86.
Lakens D, Scheel AM and Isager PM (2018) Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science* **1**(2), 259–269.
Laurer M, van Atteveldt W, Casas A and Welbers K (2023) Lowering the language barrier: Investigating deep transfer learning and machine translation for multilingual analyses of political texts. *Computational Communication Research* **5**(2).
Laurer M, Van Atteveldt W, Casas A and Welbers K (2024) Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Political Analysis* **32**(1), 84–100.
Lehmann P, Franzmann S, Burst T, Matthieß T, Regel S, Riethmüller F, Volkens A, Weßels B and Zehnter L (2023) *Manifesto project dataset*. V. 2023a.
Lehmann P and Zobel M (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 (2023) Cross-lingual classification of political texts using multilingual sentence embeddings. *Political Analysis* **31**(3), 366–379.
Licht H and Lind F (2023) Going cross-lingual: A guide to multilingual text analysis. *Computational Communication Research* **5**(2), 1–31.
Lind F, Heidenreich T, Kralj C and Boomgaarden HG (2021) Greasing the wheels for comparative communication research: Supervised text classification for multilingual corpora. *Computational Communication Research* **3**(3). 10.5117/CCR2021.3.001.LIND.
Lucas C, Nielsen RA, Roberts ME, Stewart BM, Storer A and Tingley D (2015) Computer-assisted text analysis for comparative politics. *Political Analysis* **23**(2), 254–277.
Lyu C, Du Z, Xu J, Duan Y, Wu M, Lynn T, Aji AF, Wong DF and Wang L. 2024. A paradigm shift: The future of machine translation lies with large language models. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) May 2024. Torino, Italia: ELRA/ICCL.
Mate A, Sebők M, Wordliczek L, Stolicki D and Feldmann Á (2023) Machine translation as an underrated ingredient? solving classification tasks with large language models for comparative research. *Computational Communication Research* **5**(2). https://doi.org/10.5117/CCR2023.2.6.MATE.
Plenter JI (2023) Advantages and pitfalls of machine translation for party research: The translation of party manifestos of european parties using DeepL. *Frontiers in Political Science* **5**, 1268320.
Poljak Ž (2023) Parties' attack behaviour in parliaments: Who attacks whom and when. *European Journal of Political Research* **62**(3), 903–923.
Rauh C and Schwalbach J. 2020. *The ParlSpeech v2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies*.
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.
Theocharis Y, Barberá P, Fazekas Z, Popa SA and Parnet O (2016) A bad workman blames his tweets: The consequences of citizens' uncivil twitter use when interacting with party candidates. *Journal of Communication* **66**(6), 1007–1031.
Tiedemann J and Thottingal S. 2020. OPUS-MT – Building open translation services for the world. In *Proceedings of the 22nd annual conference of the european association for machine translation*, pp. 479–480.
Vries ED, Schoonvelde M and 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.
Wang Y (2023) Topic classification for political texts with pretrained language models. *Political Analysis* **31**(4), 662–668.
Wankmüller S (2024) Introduction to neural transfer learning with transformers for social science text analysis. *Sociological Methods & Research* **53**(4), 1676–1752. doi:10.1177/0049124122113452.
Windsor LC, Cupit JG and Windsor AJ (2019) Automated content analysis across six languages. *PloS One* **14**(11), e0224425.
Zhang, T., Kishore, V., Felix, W., Weinberger, K. Q. and Artzi, Y. (2020) BERTScore: Evaluating text generation with BERT, Proceedings of the 8th International Conference on Learning Representations (ICLR), 26–30 April.

Cite this article: Licht H, Sczepanski RI, Laurer M and Bekmuratovna A (2026) Validating open-source machine translation for quantitative text analysis. *Political Science Research and Methods*, 1–19. https://doi.org/10.1017/psrm.2026.10102

https://doi.org/10.1017/psrm.2026.10102 Published online by Cambridge University Press