ECONtribute
MARKETS & PUBLIC POLICY

ECONtribute
Discussion Paper No. 276

No More Cost in Translation: Validating Open-Source Machine Translation for Quantitative Text Analysis

Hauke Licht
Moritz Laurer
Ronja Sczepanski
Ayjeren Bekmuratovna

February 2024

www.econtribute.de

UNIVERSITÄT BONN
Universität zu Köln

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1-390838866 is gratefully acknowledged.

No more cost in translation: Validating open-source machine translation for quantitative text analysis

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

[1]University of Cologne

[2]Sciences Po Paris

[3]Hugging Face

[4]Vrije Universiteit Amsterdam

[5]DHL

February 5, 2024

Abstract

As more and more scholars apply computational text analysis methods to multilingual corpora, machine translation has become an indispensable tool. However, relying on commercial services for machine translation, such as Google Translate or DeepL, limits reproducibility and can be expensive. This paper assesses the viability of a reproducible and affordable alternative: free and open-source machine translation models. We ask whether researchers who use an open-source model instead of a commercial service for machine translation would obtain substantially different measurements from their multilingual corpora. We address this question by replicating and extending an influential study by de Vries et al. (2018) on the use of machine translation in cross-lingual topic modeling, and an original study of its use in supervised text classification with Transformer-based classifiers. We find only minor differences between the measurements generated by these methods when applied to corpora translated with open-source models and commercial services, respectively. We conclude that "free" machine translation 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.

*Corresponding Author: hauke.licht@wiso.uni-koeln.de

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. 2022; cf. Baden et al. 2022, Dolinksy et al. 2022, Licht and Lind 2023). It allows bridging language barriers by transferring documents written in different languages into a single target language and thus 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).

To date, however, most scholars rely on commercial services for machine translation, such as Google Translate or DeepL (but see Licht 2023; Laurer et al. 2023; Mate et al. 2023). This approach comes with clear limitations. First, using commercial services limits reproducibility because the underlying translation models are closed-source and not versioned (Chan et al. 2020). Second, machine-translating large amounts of text can be expensive because commercial services charge users for each translated character.[1]

This paper argues for the viability of an affordable, transparent, and reproducible alternative: using open-source models for machine translation. Open-source MT models, such as OPUS-MT (Tiedemann and Thottingal 2020) or Facebook Research's M2M model (e.g. Fan et al. 2021), allow researchers to translate large text corpora without needing to pay fees to a commercial service. Moreover, using these models for machine translation ensures reproducibility because they are publicly available for download.

While open-source MT models promise cheaper and reproducible machine translation, applied researchers must know whether they enable reliable cross-lingual quantitative text analysis of political text corpora. We thus assess whether machine-translating multilingual corpora with available open-source models instead of a commercial service (Google

1. see Table 1 on page 6

Translate or DeepL) reduces reliability and yields substantially different measurements when applying computational text analysis methods. First, we extend the seminal study by de (Vries et al. 2018), who evaluate the machine translation approach for cross-lingual topic modeling (Vries et al. 2018). Second, we present an original study of the reliability of machine translation for supervised text classification with Transformer-based language models.

Our findings support the conclusion that open-source MT models can be a reliable replacement for commercial services when applying bag-of-words as well as 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 the case of 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 on speech translations generated with Google Translate. Further, both machine translation-based models allocate speeches to similar topics as the human translation-based 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 perform, on average, only 0.007 F1 score points worse than comparable classifiers fine-tuned using translations from a commercial MT model as input.

We conclude that "free" machine translation 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 machine translation in applied research, we provide an online translation application.[2]

2. The application is available online

2 Machine translation for quantitative text analysis

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

In their seminal study, Lucas et al. (2015) argue that comparative researchers can use machine translation 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 first in supporting Lucas et al.'s argument with extensive comparative evidence. They base their study on a subset of the EuropParl parallel corpus (Koehn 2005), which contains the original text of speeches held in the European Parliament and their translations into the EU's official languages 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 machine translation 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.

De Vries et al. (2018) study has been highly influential. It is frequently cited in applied research to justify a machine translation approach to cross-lingual bag-of-words text analyses (e.g., Barberá et al. 2022). Further, 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. Further, Düpont and Rachuj (2022) evaluate the machine translation approach for comparing the textual similarity of documents across languages. Courtney et al. (2020) presents evidence on the reliability of machine translation for bag-of-words supervised text classification. They examine whether supervised text classifiers trained on English-language machine

translations 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: an affordable, transparent, and reproducible alternative

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

We argue that open-source machine translation (MT) models, such as OPUS-MT (Tiedemann and Thottingal 2020) and M2M (Fan et al. 2021), offer a promising alternative due to their cost-effectiveness, reproducibility, and transparency.[3] Specifically, using open-source models is cheaper because researchers with some programming experience only need to invest in GPU computing resources instead of paying a fee for translation. Further, using open-source MT models is more transparent than a commercial service and ensures the reproducibility of analyses involving machine translation.

Cost efficiency and resource requirements

Using commercial services to machine-translate large text corpora can be expensive, as one pays for each translated character. 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

3. Please refer to the Supporting Materials, Section A.1, for details about these MT models.

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-418m	A100	2.36	1.6
18 mio.	M2M-418m	V100	2.36	3.9 [b]
18 mio.	M2M-418m	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

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-418m). The two main costs incurred by MT 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[4] costs, which are easily accessible through services like *Google Colab*.[5]

Our cost estimates lead to two important insights: First, 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. The trade-off is between compute time and expertise. Especially on older GPUs, translations 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 provide an interactive Google Colab-based online translation application,[6] code base, and tutorial. We

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

5. For commercial services, the costs are: USD 20 per 1 million characters for Google Translate and USD 20 per 1 million characters for DeepL. Google Colab makes GPUs accessible either for free (with lower reliability) or for around EUR 11 for a fixed budget of compute hours. For our our cost estimate in table below, we used a set of 500 long parliamentary speeches, 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.

6. The application is available online

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 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.

Transparency and reproducibility

While some might find the practical usability of commercial MT services appealing and the costs their use creates negligible, using commercial MT services raises concerns about reproducibility and transparency. First, 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 might 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. Researchers can thus document the exact version of the model they have used, which makes 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. Further, it is a well-established best practice to report models' performances on 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. 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.

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 comparatively reliable cross-lingual quantitative text analysis in comparison to commercial services (cf. Vries et al. 2018; Reber 2019; Courtney et al. 2020; Windsor et al. 2019).

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 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.

Our studies cover two widely used quantitative text analysis methods: topic modeling and supervised text classification. Further, our studies are based on corpora from different domains of political communication (parliamentary speech politics, party manifestos, and social media).

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 comparing the in- and outputs of a topic model obtained from machine-translated texts to those obtained from human expert translations. This strategy provides an ideal comparison because human experts are "gold-standard" translators. For both corpora, they pre-processed the text

data,[7] created Term-Document-Matrices (TDMs) that count the number of occurrences of words ("terms") in a speech ("document"), and fit 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 stems and whether the same documents have similar probabilities of being in a specific topic. Overall, they find that the topic model based on machine-translated speeches is very similar to the one fitted on human translations when it comes to the content of a topic and how documents are matched to the topics. We take advantage of this strength of their study in our analysis.

3.1.1 Empirical strategy

Like de Vries et al. (2018), we compare the reliability of machine translation 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 it shows higher discrepancies when we use an open-source MT model for translations instead of Google Translate, this would indicate that relying on open-source models for translation impairs the reliability of 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 fare 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' reliability compared to the human translation benchmark. To compare the quality of open-source machine translation, we translated the parallel corpora from the respective language into English with OPUS-MT (Tiedemann and Thottingal 2020).[8] We then pre-processed the data in the same way

7. deleting speeches containing less than 50 words, removing stop words, numbers, and punctuation, as well as stemming and lower casing words

8. We chose OPUS-MT because it had the lowest translation costs regarding time and CO^2 emissions

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 in- and outputs of topic models fitted to OPUS-MT or Google Translate translations to the ones obtained from human expert-translated texts.[9] We used the same metrics as de Vries et al. (2018): The similarity between documents' bag-of-word representations. The similarity between documents' estimated topic proportions. And the similarity between 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.

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 machine translation source affects the bag-of-words inputs to the topic model. Specifically, we compare the term-document matrices (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 Supporting Materials). Specifically, the two machine translation 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.[10] Therefore, using OPUS-MT for machine translation resulted in document representations that are as similar to those obtained from

produced by the GPU compute time.

9. To make topic models' outputs comparable, similar to Vries et al. (2018), we have matched the topics based on the stems allocated to the topics.

10. Google Translate has a slight advantage in Polish, German, and Danish, and OPUS-MT performs exceptionally well in Roman languages such as French and Spanish.

Table 2: Summary statistics of cosine similarities between bag-of-words representations' onbtained from machine- and human-translated texts at 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

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

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 substantially lower (see Table B1 in the Supporting Materials). 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, as overlap in word frequencies is also important for topic modeling tasks.

As the next step of our analysis, we assess how the machine translation 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 Figure B2 in the Supporting Materials). 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 exceptionally well for French, with an average correlation of 0.801. On the other hand, Google Translate has a higher correlation with the gold standard for German, with an average correlation of 0.809. Overall, however, 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.[11] And as is shown in Figure B3 in the Supporting Materials, 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 stems of topics in the human-translated data-based topic models.

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

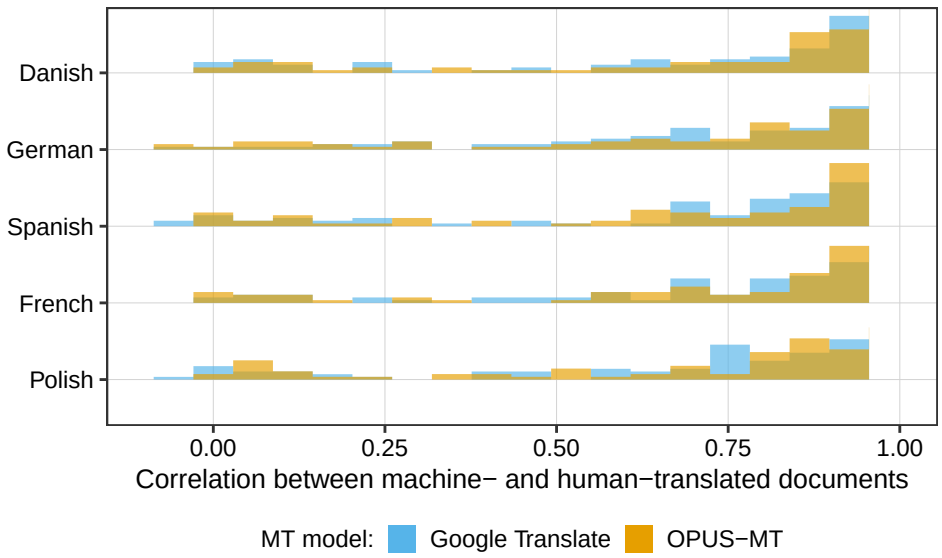

Figure 1: Similarity of corpus-level topical prevalence.

3.2 Study 2: Machine translation in supervised classification

Should researchers expect that supervised classifiers are less reliable in a labeling task when fine-tuned using texts translated with an open-source instead of a commercial MT model?

In our second study, we address this question with a large comparative experiment that allows us to estimate the expected performance difference 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 labeled multilingual political text data sets. Our benchmark includes four replication data sets (Düpont and Rachuj 2022; Lehmann and Zobel 2018; Poljak 2023; Theocharis et al. 2016). As shown in Table 4, these data sets jointly cover 10 European languages and three domains of political communication (parliamentary speech, party manifestos, and social media).

The texts in these four datasets have been coded on multiple dimensions. As shown in Table 5, our benchmark thus covers various target concepts, ranging from sentiment over negative campaigning to incivility. We use this variation within and across datasets

Table 4: Datasets

Description	Languages
Düpont & Rachuj (2022): sentences from manifestos taken from the CMP corpus	Danish, Dutch, Finish, French, German, Italian, Spanish, Swedish
Lehmann & Zobel (2018): quasi-sentences from manifestos taken from the CMP corpus	Danish, Dutch, English, Finish, 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 and Croatian
[b] additional analyses conducted, including tweets written in Greek

to define 15 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 are not specific to a single data set, target concept, or classification task.

We have obtained machine translations into English for all non-English texts in our four data sets from two commercial services (DeepL and Google Translate) and three open-source MT models (M2M 418M, M2M 1.2B, and OPUS-MT). In addition, all but the Lehmann and Zobel (2018) data come with Google Translate translations that the data sets' owners obtained when compiling them.[12]

3.2.2 Empirical strategy

Our main research question is whether it makes a difference for a classifier's reliability to use an open-source instead of a commercial MT model to translate the texts taken as inputs when fine-tuning it. Accordingly, we are not primarily interested in how well classifiers perform in a task but how classifiers fine-tuned for the same task compare when

12. 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.

Table 5: Tasks overview

Data	Task	Label classes
Düpont & Rachuj: sentences from manifestos taken from the CMP corpus		
CMP major policy domain categories	classify the policy topic discussed in quasi-sentence	extrel, freedem, polsys, econ, welqual, fabsoc, socgrp
CMP left–right position indicators	classify the stance expressed in quasi-sentence	left, right, none
	classify stance expressed in quasi-sentence (binary)	left, right
CMP left–right position indicators in domain category “Economy"	classify the stance expressed in quasi-sentences about economic issues (binary)	left, right
CMP left–right position indicators in damain category “Freedom & Democracy"	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		
PimPo issue categories	classify the issue focus of quasi-sentences	immigration, integration
PimPo position indicator	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		
dichotomized attack count indicator	detect whether a speech contains one or more attacks of parliamentary actors	attack: yes, no
attack type indicators	classify the type of attack	attack type: policy, trait, both
incivility indicator	detect incivile attacks	incivile: yes, no
Theocharis et al.: tweets, retweets, and replies to tweets by candidates for the 2014 European Parliament election		
sentiment categories	classify the sentiment of tweets	positive, neutral, negative
	classify the sentiment of tweets (binary)	positive, negative
type of communication indicator	classify the type of communication in tweets	broadcasting, engaging
politeness indicator	detect impolite tweets	polite, impolite
tweet focus indicator	detect political tweets	political: yes, no

using different translations as inputs.

We address this question empirically by fine-tuning one classifier per text transla-

tion model (i.e., one using DeepL translations, one using Google Translate translations, etc.) for each of our 15 tasks.[13] This results in six (five) classifiers per task for datasets with(out) old Google Translate translations.[14] In addition, we have fine-tuned one multilingual classifier per task to also allow for additional comparisons to the alternative strategy of aligning texts through embedding instead of translation (cf. Licht and Lind 2023).[15]

We then compare these classifiers' language- and label class-specific performances in labeling held-out texts within tasks, using the F1 score as an evaluation metric to estimate classifiers' reliability.[16] Holding constant the random seed, data splits, and fine-tuning hyper-parameters for each task,[17] we can directly compare the predictions and out-of-sample classification performances of classifiers fine-tuned using different machine translations of the same texts. We can thus put classifiers fine-tuned using texts machine-translated with commercial MT services into a head-to-head comparison with classifiers fine-tuned using machine translations generated with open-source models. This allows us to assess, for example, whether a sentiment classifier fine-tuned using OPUS-MT translations achieved a lower F1 score in classifying originally Spanish texts than a comparable classifier fine-tuned using DeepL translations.

Our main analysis focuses on estimating the average performance difference of classifiers fine-tuned using open-source MT models' translations relative to classifiers fine-tuned using commercial MT models' translations. To this end, we combine classifiers' language- and label class-specific F1 scores across tasks and regress these scores on an indicator of the (type of) MT model used to translate texts.[18] To account for uncertainty in estimates of classifiers' test set F1 scores, we use 50 bootstrapped F1 scores for each classifier as

13. 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 (see Table C6).

14. We have used the RoBERTa base checkpoint for fine-tuning for all except those in the Theocharis et al. (2016) data set. As the Theocharis et al. (2016) data set records Twitter posts, we used the RoBERTa-based Twitter language model pre-trained by Barbieri et al. (2020) instead.

15. using XLM-T (Barbieri et al. 2022) for the tasks in the Theocharis et al. (2016) data set and XLM-RoBERTa (Conneau et al. 2020) instead

16. We focus on the F1 score because the label classes are imbalanced in all our tasks (cf. Table C1–C4).

17. see Table C6

18. We exclude F1 scores in classifying English-language texts from this comparison, as they would downward bias our performance difference estimates. However, our results are robust to including them.

outcomes in our regressions. Further, to account for heterogeneity across classifiers (cf. Figures C1-C4), all our regressions include data set, 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 regression:

$$\text{F1 score}_{d,t,c,l} = \beta_0 + \beta_1 \,\text{model type} + \delta_d + \theta_t + \kappa_c + \lambda_l + \epsilon_{d,t,c,l} \tag{1}$$

where $\text{F1 score}_{d,t,c,l}$ represents a bootstrapped F1 score for dataset d, task t, label class c, and language l; "model type" is a categorical indicator differentiating between commercial MT-based classifiers (reference category), open-source MT-based translations, and multilingual Transformer-based classifiers; and δ, θ, κ, and λ are the data set, task, label class, and language fixed effects estimates, respectively.

If 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-our texts. The coefficient estimate's magnitude, in turn, indicates by how much worse.

Note that as our goal is to compare the classifiers fine-tuned for each task using the translations obtained with different MT models, we restrict our analyses to the languages that can be translated into English by all models. For example, we excluded the Greek tweets in the Theocharis et al. (2016) data set because they cannot be translated to English by the OPUS-MT model (see the note in Table 4).

3.2.3 Results

Table 6 reports the results of our main regression analyses. The evidence it presents speaks directly to our research question of what reduction in classification reliability we should expect when using an open-source MT model instead of a commercial MT service to translate the text inputs taken to fine-tune a supervised text classifier.

Our evidence suggests we should expect such a reduction, but it will be negligibly small. Model 1 estimates the average difference in classifiers' test set F1 scores when using

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
Type of model (ref.: commercial MT model)		
open-source MT model	-0.007 $(0.001)^{***}$	
multilingual classifier	-0.012 $(0.002)^{***}$	
Translation model (ref.: DeepL)		
Google Translate		0.006 $(0.002)^{**}$
Google Translate (old)		0.001 (0.002)
OPUS-MT		0.002 (0.002)
M2M (1.2B)		-0.002 (0.002)
M2M (418M)		-0.013 $(0.002)^{***}$
multilingual		-0.009 $(0.002)^{***}$
R^2	0.428	0.429
Adj. R^2	0.428	0.429
Num. obs.	48300	48300

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.
The F1 scroe is measured on a scale from 0 to 1. A coefficient estimates 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 points.
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.

an open-source MT model instead of a commercial one. This difference is estimated to be negative and statistically significant ($t = -5.42$, $p < 0.000$). However, the estimated magnitude is only 0.007, less than a difference of 0.01 units on the [0, 1] F1 score scale.

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 data set (Licht 2023; Laurer et al. 2022) or with different seed (Wang 2023) usually exhibit higher levels of variability in test set F1 scores than our estimate of 0.007.

Moreover, Model 1 in Table C8 in the Supporting Materials shows that these findings hold when dropping classifiers fine-tuned with input text translations generated with older Google Translate versions from the comparison. Model 2 in Table C8 further presents

evidence that the estimated classification reliability reduction of 0.007 in model 1 in Table 6 drops by 60% and becomes statistically insignificant 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 further 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 neither using OPUS-MT nor the 1.2B parameter M2M model instead of DeepL significantly reduces classifiers' test set F1 scores. Further, it shows that even the "worst" alternative – using the small M2M model – only reduces classifiers' test set F1 score by 0.013 compared to using DeepL. Again, we believe that this difference is practically negligible.

We visualize a more detailed breakdown of the results for Model 2 in Table 6 in Figure C5 in the Supporting Materials. It reports the results from regressions that pairwise compare translation models in how they affect classifiers' out-of-sample performance and underscores that OPUS-MT and the large M2M model are competitive alternatives to commercial machine translation services.

Moreover, in Table C10 in the Supporting Materials, we show that our main finding holds when comparing classifiers' text-level predicted labels instead of their overall reliability. Specifically, we find that the labels predicted for test set samples by classifiers fine-tuned using M2M 1.2B or OPUS-MT translations agree on average no less with the labels predicted by DeepL-based classifiers than the labels predicted by Google Translate-based classifiers. This, again, underscores that using open-source MT models results in comparable degrees of measurement reliability than using commercial MT services when fine-tuning translation-based text classifiers.

4 Conclusion and discussion

Open-source machine translation (MT) models like OPUS-MT (Tiedemann and Thottingal 2020) and M2M (Fan et al. 2021) are affordable, transparent, and reproducible alter-

natives to commercial MT services like Google Translate and DeepL. We have assessed whether machine-translating multilingual corpora with available open-source models instead of a commercial service (Google Translate) yields substantially different results when applying two common computational text analysis methods. Our first study replicates and extends the study by Vries et al. (2018), who evaluate machine translation for cross-lingual topic modeling. Our second study is an original analysis of the reliability of open-source machine translation for cross-lingual supervised text classification with Transformer-based classifiers.

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. Further, both machine translation-based models allocate speeches to similar topics as 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 perform, on average, only 0.7 F1 scores worse than comparable classifiers fine-tuned using translations from a commercial MT model as input.

Our findings has important implications for applied researchers. Given that using "free" MT models for topic modeling or fine-tuning a Transformer-based classifier results in no less reliability measurements than using a commercial MT service, applied researchers can benefit from the transparency and reproducibility advantage. Maybe as important from a practical point of view, using open-source MT models can save researchers costs. As a point in case, 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 U.S. $ 1267 (see Table C5). And the labeled texts in our four benchmark data sets make up only a fraction of the target corpora analyzed in the papers they originate from.

Our study is not without limitations, however. Researchers might find translation with open-source MT models technically challenging. While software packages such as the `EasyNMT` Python package provide a handy toolkit,[19] we acknowledge that deploying these models and using GPU computing environments are no trivial skills. The code base and online app we provide thus aim to lower this accessibility barrier.[20]

Further, 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 the language other languages descended from. Our study does not provide evidence on the reliability of these alternative strategies.

Additionally, 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 large language models for prompt-based zero- and few-shot political text classification (e.g., Gilardi et al. 2023). However, given the influence and popularity of the LDA topic model and Transformer encoder fine-tuning in applied political science research, we believe that our study should still inform the methodological choices of many researchers.

Acknowledgments

This project has received funding through the *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1 – 390838866.

19. see https://easynmt.net/demo
20. see https://colab.research.google.com/drive/1quRuHPzXMIrXpMmWvUNj83knhbx8aZLF

References

Baden, C., C. Pipal, M. Schoonvelde, and M. A. C. G. van der Velden. 2022. "Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda." *Communication Methods and Measures* 16 (1): 1–8. DOI: 10.1080/19312458.2021.2015574.

Barberá, P., A. R. Gohdes, E. Iakhnis, and T. Zeitzoff. 2022. "Distract and Divert: How World Leaders Use Social Media During Contentious Politics." *The International Journal of Press/Politics,* 19401612221102030. DOI: 10.1177/19401612221102030.

Barbieri, F., J. Camacho-Collados, L. Espinosa Anke, and L. Neves. 2020. "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification." In *Findings of the Association for Computational Linguistics: EMNLP 2020,* edited by T. Cohn, Y. He, and Y. Liu, 1644–1650. Findings 2020. Association for Computational Linguistics. DOI: 10.18653/v1/2020.findings-emnlp.148.

Barbieri, F., L. Espinosa Anke, and J. Camacho-Collados. 2022. "XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* edited by N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, 258–266. LREC 2022. European Language Resources Association.

Baum, M. A., and Y. M. Zhukov. 2019. "Media Ownership and News Coverage of International Conflict." *Political Communication* 36 (1): 36–63. DOI: 10.1080/10584609.2018.1483606.

Blei, D. M., A. Y. Ng, M. I. Jordan, and J. Lafferty. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (4): 993–1022.

Chan, C.-H., J. Zeng, H. Wessler, M. Jungblut, K. Welbers, J. W. Bajjalieh, W. van Atteveldt, and S. L. Althaus. 2020. "Reproducible Extraction of Cross-Lingual Topics (rectr)." *Communication Methods and Measures* 14 (4): 285–305. DOI: 10.1080/19312458.2020.1812555.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. "Unsupervised Cross-lingual Representation Learning at Scale." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* edited by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, 8440–8451. ACL 2020. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.747.

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. DOI: 10.1080/19331681.2020.1731245.

Dancygier, R., and Y. Margalit. 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. DOI: 10.1177/0010414019858936.

Dolinksy, A., M. Schoonvelde, and M. A. C. G. van der Velden. 2022. "Multilingualism in Computational Text Analysis Methods: Evidence From A Pre-Registered Survey Experiment." In *11th Meeting of the European Political Science Association,* EPSA 2022.

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. DOI: 10.1017/S0007123420000617.

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

Gilardi, F., M. Alizadeh, and M. Kubli. 2023. "ChatGPT outperforms crowd workers for text-annotation tasks." *Proceedings of the National Academy of Sciences* 120 (30): e2305016120. DOI: 10.1073/pnas.2305016120.

Koehn, P. 2005. "Europarl: A parallel corpus for statistical machine translation," 5:79–86. Citeseer.

Laurer, M., W. v. Atteveldt, A. Casas, and K. Welbers. 2022. "Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI."

———. 2023. "Lowering the Language Barrier: Investigating Deep Transfer Learning and Machine Translation for Multilingual Analyses of Political Texts." *Computational Communication Research* 5 (2). DOI: 10.5117/CCR2023.2.7.LAUR.

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. DOI: 10.1111/1475-6765.12266.

Licht, H. 2023. "Cross-lingual classification of political texts using multilingual sentence embeddings." *Political Analysis* 0 (0): 1–14. DOI: 10.1017/pan.2022.29.

Licht, H., and F. Lind. 2023. "Going cross-lingual: A guide to multilingual text analysis." *Computational Communication Research* 5 (2): 1.

Lind, F., T. Heidenreich, C. Kralj, and H. G. Boomgaarden. 2021. "Greasing the wheels for comparative communication research: Supervised text classification for multilingual corpora." *Computational Communication Research* 3 (3). DOI: 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. DOI: 10.1093/pan/mpu019.

Mate, A., M. Sebők, L. Wordliczek, D. Stolicki, and Á. Feldmann. 2023. "Machine Translation as an Underrated Ingredient? Solving Classification Tasks with Large Language Models for Comparative Research." *Computational Communication Research* 5 (2). DOI: 10.5117/CCR2023.2.6.MATE.

Poljak, Ž. 2023. "Parties' attack behaviour in parliaments: Who attacks whom and when." *European Journal of Political Research* 62 (3): 903–923. DOI: 10.1111/1475-6765.12542.

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. DOI: 10.1080/19312458.2018.1555798.

Theocharis, Y., P. Barberá, Z. Fazekas, S. A. Popa, and O. Parnet. 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. DOI: 10.1111/jcom.12259.

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

Vries, E. de, 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. DOI: 10.1017/pan.2018.26.

Wang, Y. 2023. "Topic Classification for Political Texts with Pretrained Language Models." *Political Analysis* 31 (4): 662–668. DOI: 10.1017/pan.2023.3.

Windsor, L. C., J. G. Cupit, and A. J. Windsor. 2019. "Automated content analysis across six languages." *PloS One* 14 (11): e0224425. DOI: 10.1371/journal.pone.0224425.

Supporting Materials

No more cost in translation: Validating open-source machine translation for quantitative text analysis

A Machine translation

A.1 Open-source MT models

OPUS-MT models are small and specialized encoder-decoder Transformer models (Tiedemann and Thottingal 2020). Each model can only translate exactly one language direction (e.g. Chinese to English, but not English to Chinese). This gives individual models good performance for translating between two languages in one direction with relatively small size (around 300 MB), but it this also means that every language pair and translation direction requires a separate model. More than 1000 models are available open-source for many different directions.[21] The models are trained on the OPUS corpus, which is an open-source collection of manually translated text pairs. Moreover, data augmentation techniques such as back-translation are used to increase diversity. The models are funded by European Union and Finish grants.

The M2M model is a large and general encoder-decoder Transformer model (Fan et al. 2021). It is general because it can translate in any direction between 100 languages (9900 directions) simultaneously. This makes the model more general, but it is also significantly larger and slower than the OPUS-MT models. It exists in three sizes: 0.418, 1.2, or 12 billion parameters with large model files (1.9, 5, or 47 GB). The training data is created by automatically mining highly semantically similar texts in different languages. Monolingual texts are embedded with a multilingual embedding model and highly similar texts are then matched as probable translations. Moreover, data augmentation strategies

21. see https://huggingface.co/Helsinki-NLP and https://github.com/UKPLab/EasyNMT

Table A1: Overview of published political and communication science articles evaluating machine translation for different bag-of-words quantitative text analysis tasks.

Reference	Task	Domain	Translation service	Source language(s)	Target lang.
Lucas et al. (2015)	Topic modeling (STM)	Citizen-produced social media	Google Translate	Arabic, Chinese	English
De Vries et al. (2018)	Topic modeling (LDA)	Parliamentary speech	Expert translations, Google Translate	Danish, French, German, Spanish, Polish	English
Reber (2019)	Topic modeling (LDA)	Web pages of (I)NGOs	Google Translate, DeepL	German	English
Windsor et al. (2019)	Dictionary analysis (LWIC)	UN plenary speeches	Google Translate	English, French, German, Russian, Chinese, Arabic	English
Düpont and Rachuj (2022)	Textual similarity	Party manifestos	Google Translate	12 languages [a]	English
Courtney et al. (2020)	Supervised classification	News article paragraphs	Google Translate	German, Spanish	English
Lind et al. (2021)	Supervised classification	News articles	Google Translate	German, Hungarian, Polish, Romanian, Spanish, Swedish	English
Licht (2023)	Supervised classification	Party manifestos	M2M (Fan et al. 2021)	12 languages [b]	English
Laurer et al. (2023)	Supervised classification	Party manifestos	M2M	seven languages [c]	English
Mate et al. (2023)	Supervised classification	Parliamentary Speech	OPUS-MT (Tiedemann and Thottingal 2020)	Hungarian, Polish	English

[a] Catalan, Danish, Dutch, Finnish, French, Galician, German, Italian, Norwegian, Portuguese, Spanish, and Swedish

[b] same as Düpont and Rachuj (2022) by pooling their and data by Lehmann and Zobel (2018)

[c] English, French, German, Korean, Russian, Spanish, and Turkish

like back-translation and quality filtering techniques are used. The authors intentionally avoid English-centric decisions during model design, for example when choosing anchor languages for sentence pair mining. The model was created and open-sourced by Facebook AI.

B Supporting results for Study 1

Table B1: Number of tokens in topic models' vocabulary and share of tokens in machine translation-based topic models that overlap with tokens in vocabulary of topic model fitted to human experts' translations.

	N tokens			Token overlap	
Language	Experts	Google Translate	OPUS-MT	Google Translate	OPUS-MT
Danish	37761	37232	31472	0.754	0.707
German	36390	36575	31945	0.763	0.717
Spanish	38014	40524	32554	0.753	0.713
French	38064	33013	32790	0.741	0.712
Polish	37893	39184	32309	0.712	0.664

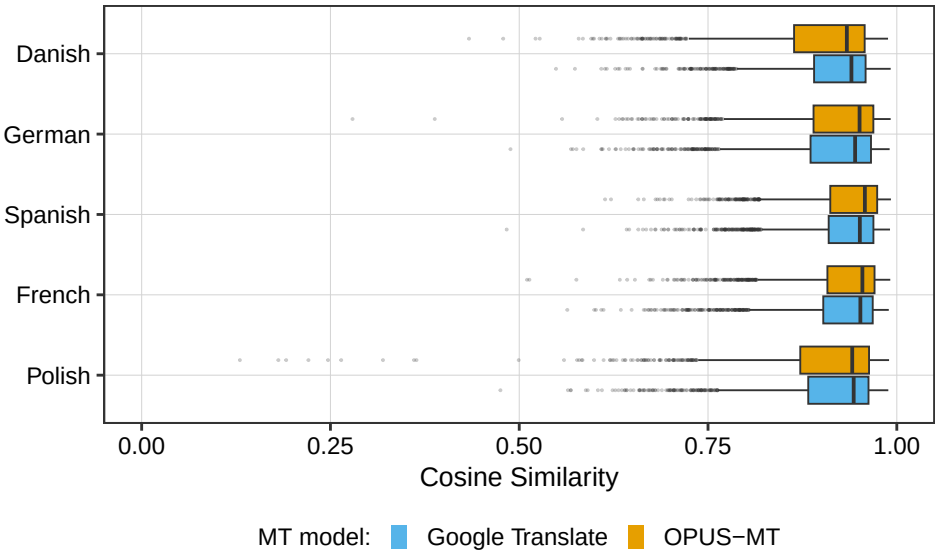

Figure B1: Distribution of cosine similarities between bag-of-words representations' onbtained from machine- and human-translated texts at document level.

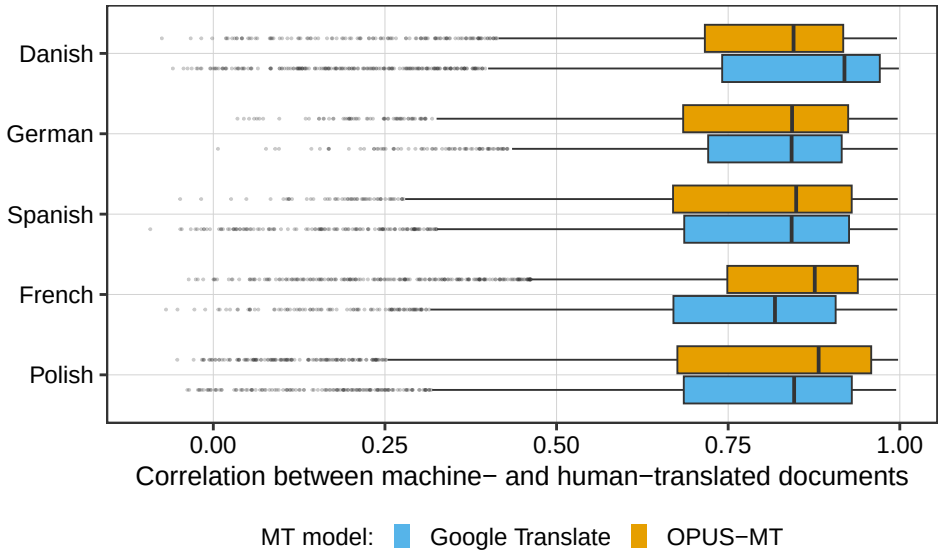

Figure B2: Similarity of document-level topic proportion estimates.

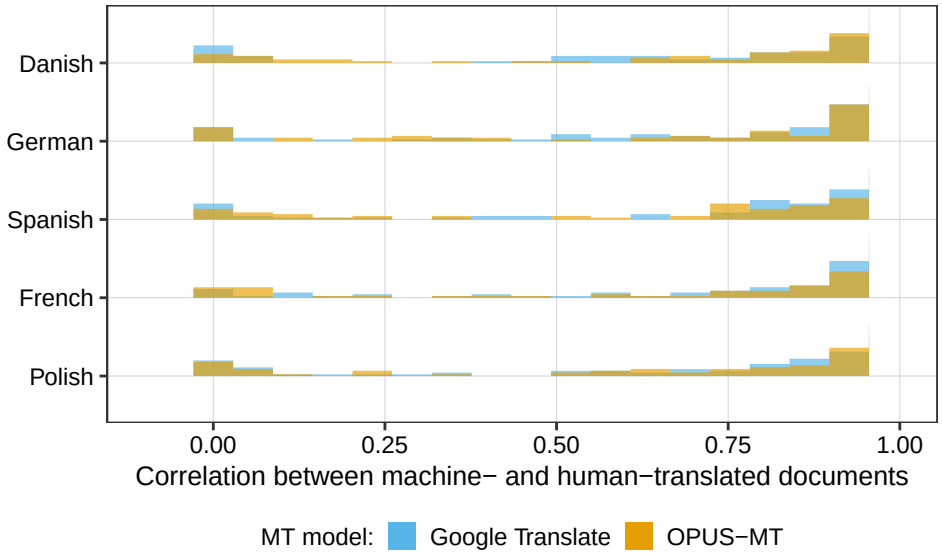

Figure B3: Comparisong of estimated topics' content between models based on human- and machine-translated texts.

C Supporting materials for Study 2

C.1 Data set descriptive statistics

Table C1: Label distribution by task and language in Düpont & Rachuj (2023) data

Task	Label class	*N*	Danish	Finish	French	Italian	Dutch	Spanish	Swedish
classify the policy topic discussed in quasi-sentence	econ	8116	0.203	0.162	0.234	0.125	0.219	0.308	0.142
	extrel	2434	0.060	0.102	0.059	0.104	0.094	0.055	0.038
	fabsoc	4409	0.049	0.205	0.109	0.093	0.228	0.047	0.142
	freedem	2257	0.032	0.069	0.075	0.050	0.056	0.080	0.059
	polsys	4249	0.061	0.095	0.219	0.071	0.093	0.138	0.111
	socgrp	2955	0.112	0.093	0.095	0.075	0.087	0.081	0.102
	welqual	9130	0.483	0.275	0.209	0.482	0.224	0.291	0.406
classify the stance expressed in quasi-sentence	left	7636	0.153	0.267	0.283	0.135	0.347	0.090	0.330
	none	16684	0.373	0.413	0.446	0.522	0.456	0.603	0.361
	right	9230	0.474	0.320	0.271	0.342	0.197	0.307	0.310
classify the stance expressed in quasi-sentences about economic issues (binary)	left	1473	0.268	0.184	0.286	0.291	0.268	0.072	0.394
	none	4860	0.520	0.490	0.384	0.505	0.593	0.700	0.497
	right	1783	0.212	0.327	0.330	0.204	0.139	0.229	0.109
classify the stance expressed in quasi-sentences about the issue of freedom and democracy (binary)	left	1065	0.643	0.150	0.261	0.854	0.717	0.415	0.838
	right	1103	0.357	0.850	0.716	0.146	0.250	0.521	0.162
	none	89			0.024		0.034	0.064	

Table C2: Label distribution by task and language in Lehmann & Zobel (2018) data

Task	Label class	*N*	Danish	English	Finish	French	Dutch	Spanish	Swedish
classify the issue focus of quasi-sentences	immigration	1963	0.512	0.603	0.381	0.69	0.399	0.520	0.233
	integration	2317	0.488	0.397	0.619	0.31	0.601	0.480	0.767
classify the stance expressed in quasi-sentences about the issues of immigration and integration (binary)	neutral	602	0.201	0.138	0.033	0.27	0.112	0.156	0.155
	sceptical	1161	0.315	0.219	0.311	0.16	0.391	0.165	0.123
	supportive	2517	0.484	0.643	0.656	0.57	0.497	0.679	0.722

C.2 Classifier fine-tuning

Table C3: Label distribution by task and language in Poljak (2022) data

Task	Label class	*N*	Bosnian	English	French	Croatian	Dutch
detect whether a speech contains one or more attacks of parliamentary actors	no	18059	0.908	0.700	0.679	0.765	0.648
	yes	6587	0.092	0.300	0.321	0.235	0.352
classify the type of attack	policy	3206	0.418	0.516	0.602	0.363	0.548
	both (policy & trait)	1752	0.272	0.203	0.226	0.346	0.293
	trait	1629	0.310	0.281	0.172	0.291	0.159
detect incivile attacks	no	5234	0.837	0.823	0.786	0.812	0.712
	yes	1353	0.163	0.177	0.214	0.188	0.288

Table C4: Label distribution by task and language in Poljak (2022) data

Task	Label class	*N*	German	Greek	English	Spanish
classify the sentiment of tweets	negative	5053	0.183	0.330	0.212	0.131
	neutral	15513	0.712	0.651	0.590	0.678
	positive	3122	0.105	0.019	0.198	0.191
classify the type of communication in tweets	broadcasting	9346	0.516	0.514	0.244	0.337
	engaging	14342	0.484	0.486	0.756	0.663
detect impolite tweets	no	2304	0.086	0.235	0.054	0.023
	yes	21384	0.914	0.765	0.946	0.977
detect political tweets	no	5987	0.449	0.136	0.268	0.169
	yes	17701	0.551	0.864	0.732	0.831

Table C5: Number of characters and estimated translation cost by language and data set

Language	Speeches	Sentences	Characters	Cost (U.S. $)
Düpont & Rachuj: sentences from manifestos taken from the CMP corpus				
Danish		1326	125034	2.50
Dutch		10734	1194246	23.88
Finish		2724	340954	6.82
French		5057	828657	16.57
Italian		827	147978	2.96
Spanish		11720	2311085	46.22
Swedish		1162	106252	2.13
Lehmann & Zobel: quasi-sentences from manifestos taken from the CMP corpus				
Danish		603	52220	1.04
Dutch		1383	150481	3.01
English		552	71900	
Finish		270	33893	0.68
French		100	18693	0.37
Spanish		965	169500	3.39
Swedish		407	30192	0.60
Poljak: parliamentary speeches delivered in Question Time sessions				
Bosnian	2606	11148	1013809	20.28
Croatian	7865	62996	7167369	143.35
Dutch	3572	51925	4919180	98.38
English	7936	45156	3856746	
French	2667	33316	3531428	70.63
Theocharis et al.: tweets, retweets, and replies to tweets by candidates for the 2014 European Parliament election				
English		6606	645000	
German		5432	540027	10.80
Greek		5703	572831	11.46
Spanish		5947	645430	12.91

Table C6: Number of epochs, training batch size, and gradient accumulation steps applied when fine-tuning our classifiers. We have held other hyper-parameters constant, using a learning rate of $1e^{-5}$, a warm-up ratio of 0.05, and a weight decay of 0.1.

Task	Epochs	Training batch size	Gradient accumulation	Downsampling ratio
Dupont & Rachuj				
classify the policy topic discussed in quasi-sentence	2	16	2	0.4
classify the stance expressed in quasi-sentence	3	16	2	0.4
classify stance expressed in quasi-sentence (binary)	3	16	2	0.4
classify the stance expressed in quasi-sentences about economic issues (binary)	5	16	2	0.4
classify the stance expressed in quasi-sentences about the issue of freedom and democracy (binary)	5	16	2	0.4
Lehmann & Zobel				
classify the issue focus of quasi-sentences	5	16	2	0.4
classify the stance expressed in quasi-sentences about the issues of immigration and integration (binary)	5	16	2	0.4
Poljak				
detect whether a speech contains one or more attacks of parliamentary actors	5	16	2	0.4
classify the type of attack	8	16	2	0.4
detect incivile attacks	8	16	2	
Theocharis et al.				
classify the sentiment of tweets	5	32	2	0.4
classify the sentiment of tweets (binary)	5	32	2	0.4
classify the type of communication in tweets	5	32	2	0.4
detect impolite tweets	5	32	2	
detect political tweets	5	32	2	

Table C7: Overall (cross-language) F1 scores by dataset, outcome, and translation model Values (in brackets) report average (95% confidence interval) of bootstrapped test set estimates. The last column reports these scores for multilingual classifiers for comparison.

	DeepL	Google Translate	Google Translate (old)	OPUS-MT	M2M (1.2B)	M2M (418M)	multilingual
Dupont & Rachuj							
classify the policy topic discussed in quasi-sentence	0.583 [0.576, 0.591]	0.589 [0.582, 0.594]	0.587 [0.579, 0.594]	0.588 [0.581, 0.597]	0.580 [0.572, 0.587]	0.573 [0.565, 0.581]	0.573 [0.564, 0.581]
classify the stance expressed in quasi-sentence	0.637 [0.630, 0.646]	0.645 [0.636, 0.653]	0.640 [0.632, 0.649]	0.634 [0.626, 0.642]	0.636 [0.629, 0.644]	0.625 [0.615, 0.632]	0.629 [0.622, 0.637]
classify stance expressed in quasi-sentence (binary)	0.754 [0.743, 0.769]	0.759 [0.741, 0.773]	0.747 [0.734, 0.757]	0.756 [0.742, 0.770]	0.756 [0.743, 0.770]	0.736 [0.724, 0.749]	0.757 [0.743, 0.768]
classify the stance expressed in quasi-sentences about economic issues (binary)	0.716 [0.689, 0.738]	0.698 [0.671, 0.724]	0.718 [0.696, 0.737]	0.698 [0.672, 0.721]	0.728 [0.696, 0.757]	0.700 [0.675, 0.722]	0.711 [0.682, 0.735]
classify the stance expressed in quasi-sentences about the issue of freedom and democracy (binary)	0.864 [0.844, 0.891]	0.865 [0.845, 0.889]	0.864 [0.846, 0.890]	0.860 [0.836, 0.889]	0.856 [0.834, 0.883]	0.851 [0.829, 0.876]	0.884 [0.867, 0.908]
Lehmann & Zobel							
classify the issue focus of quasi-sentences	0.869 [0.856, 0.883]	0.856 [0.841, 0.871]		0.863 [0.850, 0.878]	0.861 [0.847, 0.881]	0.851 [0.834, 0.870]	0.851 [0.834, 0.873]

classify the stance expressed in quasi-sentences about the issues of immigration and integration (binary)	0.877 [0.865, 0.891]	0.872 [0.859, 0.885]		0.869 [0.855, 0.883]	0.879 [0.863, 0.892]	0.868 [0.854, 0.881]	0.855 [0.842, 0.869]
Poljak							
detect whether a speech contains one or more attacks of parliamentary actors	0.786 [0.767, 0.797]	0.783 [0.765, 0.803]	0.775 [0.759, 0.792]	0.782 [0.761, 0.799]	0.774 [0.755, 0.789]	0.775 [0.758, 0.794]	0.770 [0.749, 0.792]
classify the type of attack	0.622 [0.587, 0.655]	0.621 [0.590, 0.663]	0.616 [0.578, 0.659]	0.623 [0.578, 0.661]	0.601 [0.558, 0.632]	0.622 [0.589, 0.652]	0.596 [0.569, 0.627]
detect incivile attacks	0.557 [0.503, 0.610]	0.560 [0.514, 0.614]	0.541 [0.483, 0.584]	0.550 [0.501, 0.606]	0.546 [0.503, 0.584]	0.548 [0.501, 0.597]	0.503 [0.460, 0.542]
Theocharis et al.							
classify the sentiment of tweets	0.689 [0.671, 0.702]	0.691 [0.673, 0.706]		0.688 [0.667, 0.701]	0.696 [0.677, 0.709]	0.664 [0.647, 0.678]	0.690 [0.677, 0.703]
classify the sentiment of tweets (binary)	0.929 [0.908, 0.946]	0.926 [0.909, 0.944]		0.923 [0.906, 0.937]	0.913 [0.896, 0.932]	0.912 [0.898, 0.927]	0.922 [0.905, 0.938]
classify the type of communication in tweets	0.810 [0.799, 0.826]	0.815 [0.801, 0.828]		0.818 [0.806, 0.832]	0.809 [0.794, 0.821]	0.807 [0.795, 0.820]	0.840 [0.830, 0.850]
detect impolite tweets	0.374 [0.332, 0.418]	0.376 [0.328, 0.424]		0.374 [0.336, 0.410]	0.348 [0.301, 0.396]	0.353 [0.304, 0.404]	0.430 [0.379, 0.475]
detect political tweets	0.763 [0.743, 0.786]	0.763 [0.744, 0.782]		0.756 [0.739, 0.776]	0.745 [0.723, 0.765]	0.738 [0.723, 0.759]	0.769 [0.750, 0.784]

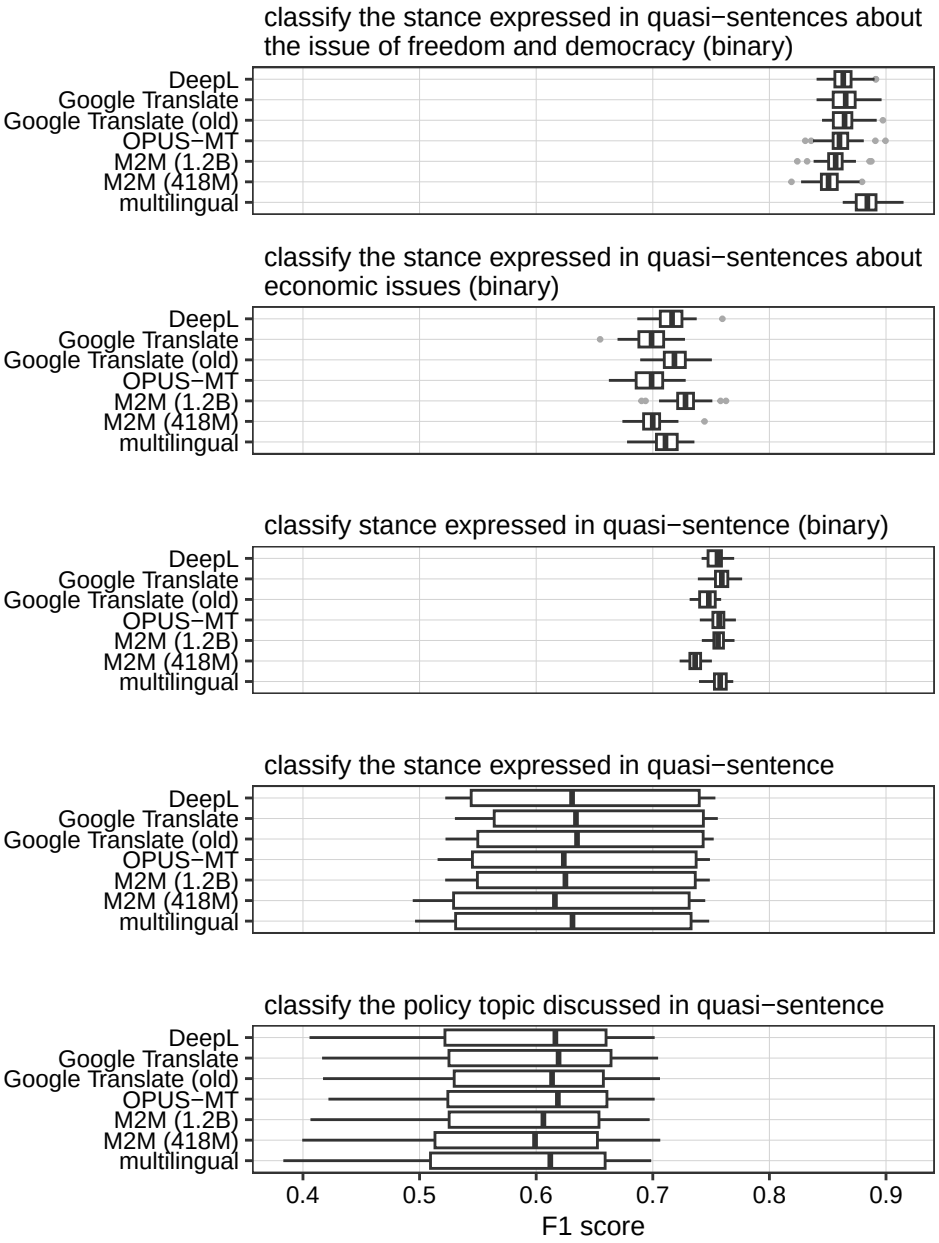

Figure C1: Summary of fine-tuned classifiers' language-specific (macro) F1 scores by task (panels) and translation source (y-axis) for Dupont & Rachuj (2022) data

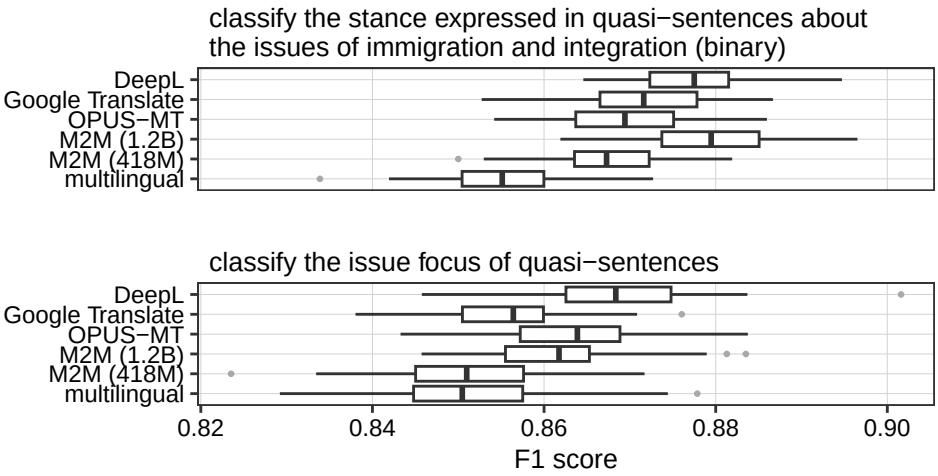

Figure C2: Summary of fine-tuned classifiers' language-specific (macro) F1 scores by task (panels) and translation source (y-axis) for Lehmann & Zobel (2018) data

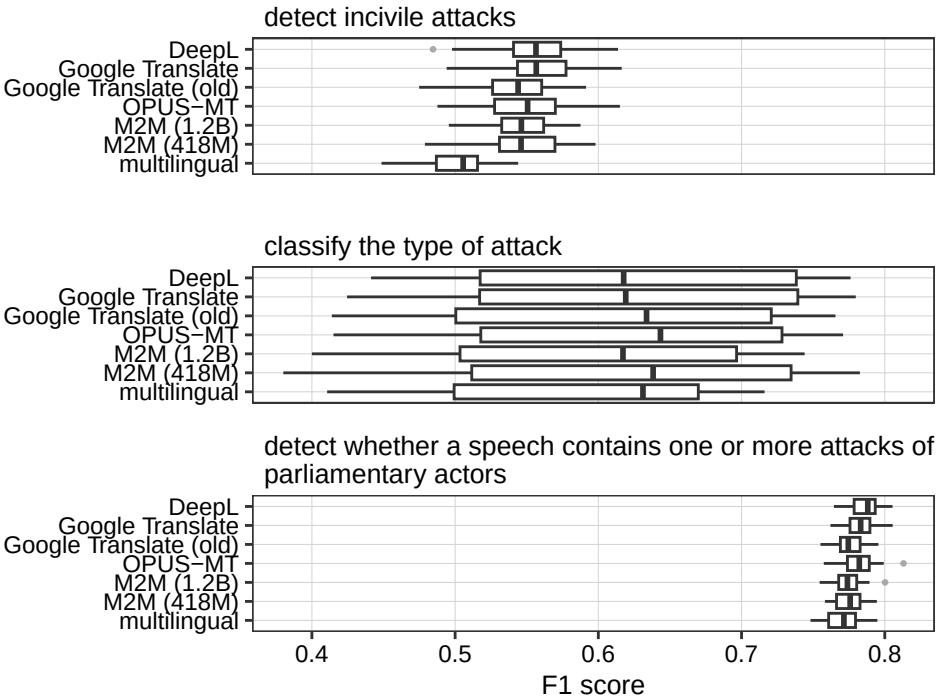

Figure C3: Summary of fine-tuned classifiers' language-specific (macro) F1 scores by task (panels) and translation source (y-axis) for Poljak (2023) data

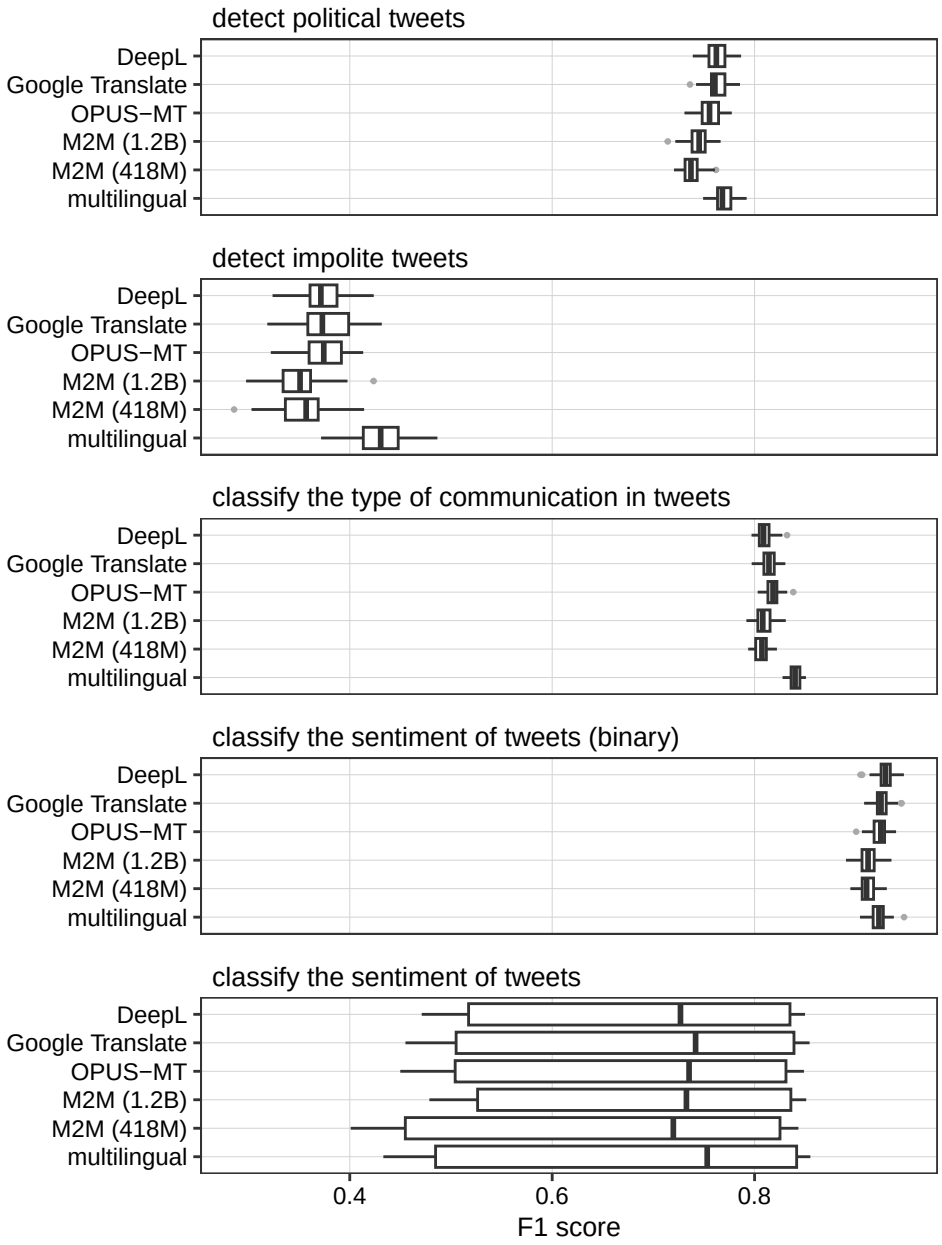

Figure C4: Summary of fine-tuned classifiers' language-specific (macro) F1 scores by task (panels) and translation source (y-axis) for Theocharis et al. (2016) data

Table C8: Additional analyses of effect of using open-source vs. commercial machine translation models for translating input texts on classifiers' language-specific out-of-sample classification performance (F1 scores). Model 1: w/o old GT translations. Model 2: w/o small M2M translations.

	Model 1	Model 2
Type of model (ref.: commercial MT model)		
open-source MT model	-0.007 $(0.001)^{***}$	-0.003 (0.001)
multilingual classifier	-0.012 $(0.002)^{***}$	-0.012 $(0.002)^{***}$
R^2	0.439	0.430
Adj. R^2	0.438	0.429
Num. obs.	42600	41200

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.
The F1 scroe is measured on a scale from 0 to 1. A coefficient estimates 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 points.
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.

C.3 Additional analyses

Table C9: Effect of using open-source vs. commercial machine translation models for translating input texts on classifiers' language-specific out-of-sample classification performance (F1 scores) in social media vs. other domains. Classifiers fine-tuned on old Google translations not included in comparison.

	Model 1	Model 2
Type of model (ref.: commercial MT model)		
open-source MT model	-0.007 $(0.001)^{***}$	
multilingual classifier	-0.014 $(0.002)^{***}$	
Translation model (ref.: DeepL)		
Google Translate		0.006 $(0.002)^{**}$
OPUS-MT		0.001 (0.002)
M2M (1.2B)		-0.001 (0.002)
M2M (418M)		-0.011 $(0.002)^{***}$
multilingual		-0.011 $(0.002)^{***}$
Social media vs. other domains		
social media data	0.181 $(0.004)^{***}$	0.182 $(0.005)^{***}$
open-source MT model X social media data	-0.009 (0.005)	
multilingual classifier X social media data	0.020 $(0.007)^{**}$	
Google Translate X social media data		-0.003 (0.007)
OPUS-MT X social media data		0.001 (0.007)
M2M (1.2B) X social media data		-0.014 $(0.007)^{*}$
M2M (418M) X social media data		-0.018 $(0.008)^{*}$
multilingual X social media data		0.019 $(0.008)^{*}$
R^2	0.439	0.440
Adj. R^2	0.439	0.439
Num. obs.	42600	42600

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.
The F1 scroe is measured on a scale from 0 to 1. A coefficient estimates 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 points.
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.

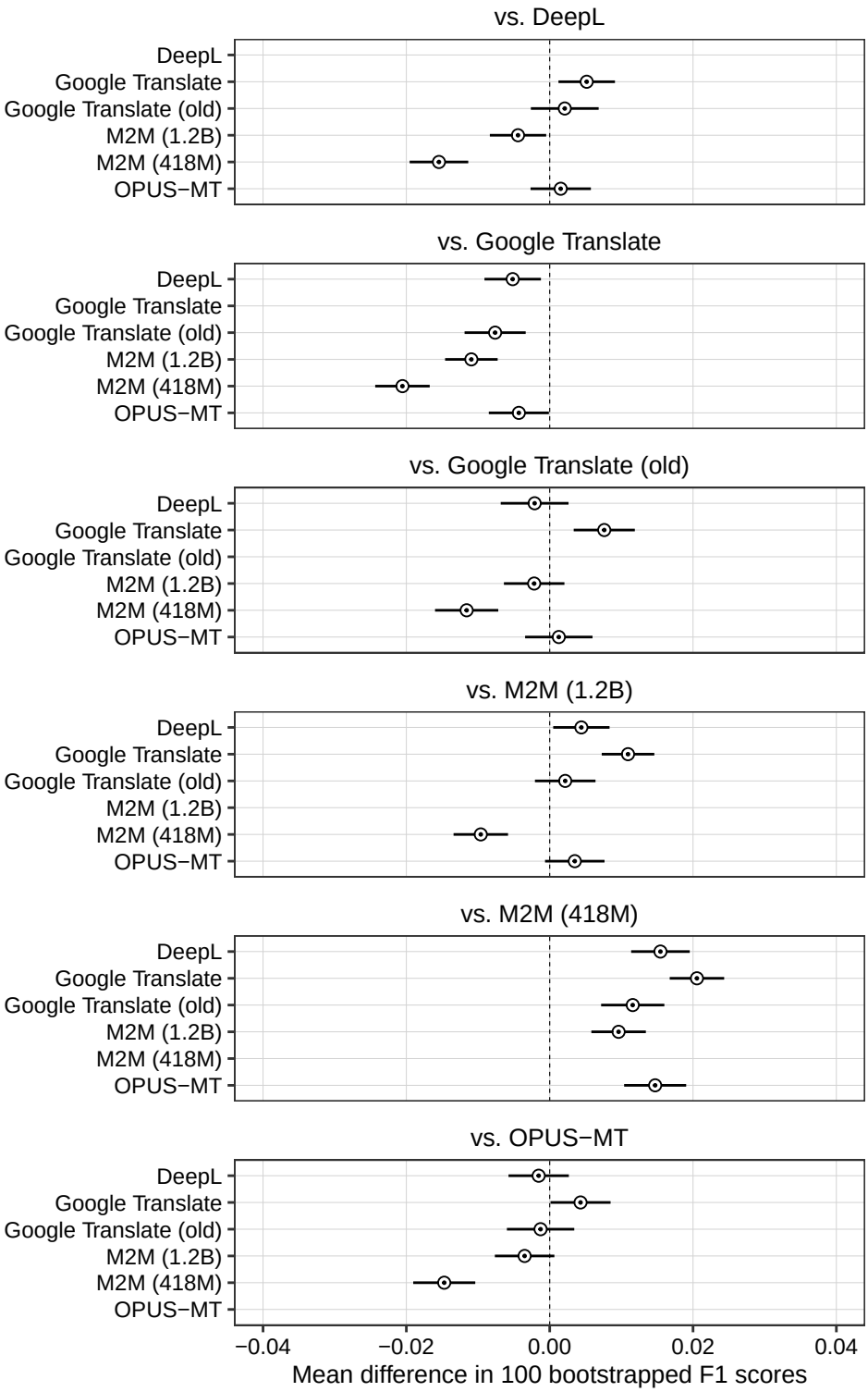

Figure C5: Summary of mean differences estimated by from regressions that compare the performances of classifiers fine-tuned using texts' translations generated with different machine translation models as input. Points (horizontal lines) indicate the mean difference (95% confidence interval) in F1 scores of the Translation model named on the y-axis compared to the translation model named in the plot panels header. For example, the positive difference for the comparison "Google Translate vs. Deepl" indicates that using DeepL instead of Google Translate to translate input texts results in, on average, more reliably classifiers.

C.3.1 Evaluation at the level of predicted levels

In addition to evaluating classifiers in terms of their overall test set performance, we can also leverage our data to compare their classifications at the level of individual texts in the test sets. This allows for a harder test of our claim that using an open-source MT model to translate the texts used to fine-tune a supervised classifier yields comparable results to using a commercial MT service.

To assess the similarity of classifiers' measurements at the level of test set predicted labels, we compare predicted labels between classifiers fine-tuned for the same task but with translations from different MT models. Because we lack human translations for our labeled data sets, we use the classifiers using DeepL translations as a benchmark in these comparisons. We then compare the agreement test set predicted labels by a classifier fine-tuned, for example, using OPUS-MT translations, with those generated by its DeepL-based counterpart. Specifically, for each task and language in the given dataset, we compute ne agreement score with the classifier based on Deepl translations for the Google Translate, OPUS-MT, M2M 418M, and M2M 1.2B-based classifiers, respectively.

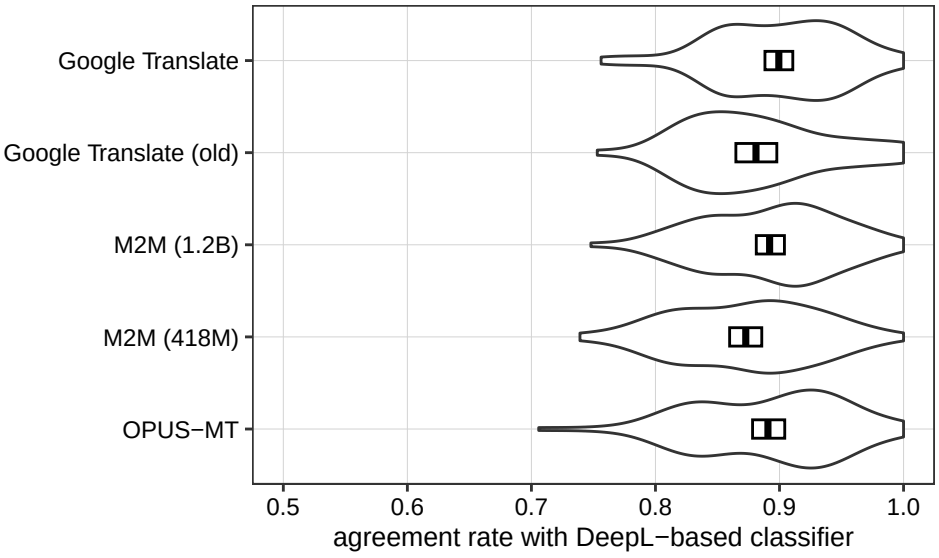

Figure C6: Distribution of classifiers' agreement with DeepL-based classifiers predicted test set labels.

Figure C6 summarizes these agreement estimates. This shows that there are no notable differences between DeepL-based classifiers predicted labels and those of classifiers

Table C10: OLS coefficient estimates of the effect of using open-source MT model instead of Google Translate on classifiers' agreement on test set examples' predicted labels realteive to DeepL-based classifiers. Google Translate-based classifiers average agreement with DeepL-based classifiers (within dataset, task, and language), shown in the intercept, used as comparison.

	Model 1
Intercept	0.856 (0.009)***
OPUS-MT	−0.011 (0.006)
M2M (1.2B)	−0.010 (0.006)
M2M (418M)	−0.030 (0.006)***
R^2	0.611
Adj. R^2	0.574
Num. obs.	280

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.
Agreement is measured on a scale from 0 to 1.
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.

fine-tuned using the other MT models' translations as inputs.

The regression results presented in Table C10 support this conclusion.[22] The estimate for the intercept of this regression model reports Google Translate-based classifiers' average agreement with their DeepL-based counterparts. The estimates for OPUS-MT, M2M 418M, and M2M 1.2B, respectively, report the average deviation of classifiers' agreement rates with DeepL-based classifiers from this baseline. Except for the coefficient for M2M 418M-based classifiers, these estimates are all statistically insignificant. This indicates that there are no systematic differences between the degrees to which the predictions of classifiers fine-tuned with these MT models translations agree with DeepL-based classifiers compared to classifiers using Google Translate translations. This underscores that the large M2M model and OPUS-MT are suitable replacements for their commercial alternatives.

22. As previously, the OLS regression includes data set, task, and language fixed effects and clusters standard errors by data set, task, and language.

References

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. DOI: 10.1080/19331681.2020.1731245.

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. DOI: 10.1017/S0007123420000617.

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

Laurer, M., W. v. Atteveldt, A. Casas, and K. Welbers. 2023. "Lowering the Language Barrier: Investigating Deep Transfer Learning and Machine Translation for Multilingual Analyses of Political Texts." *Computational Communication Research* 5 (2). DOI: 10.5117/CCR2023.2.7.LAUR.

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. DOI: 10.1111/1475-6765.12266.

Licht, H. 2023. "Cross-lingual classification of political texts using multilingual sentence embeddings." *Political Analysis* 0 (0): 1–14. DOI: 10.1017/pan.2022.29.

Lind, F., T. Heidenreich, C. Kralj, and H. G. Boomgaarden. 2021. "Greasing the wheels for comparative communication research: Supervised text classification for multilingual corpora." *Computational Communication Research* 3 (3). DOI: 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. DOI: 10.1093/pan/mpu019.

Mate, A., M. Sebők, L. Wordliczek, D. Stolicki, and Á. Feldmann. 2023. "Machine Translation as an Underrated Ingredient? Solving Classification Tasks with Large Language Models for Comparative Research." *Computational Communication Research* 5 (2). DOI: 10.5117/CCR2023.2.6.MATE.

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. DOI: 10.1080/19312458.2018.1555798.

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

Vries, E. de, 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. DOI: 10.1017/pan.2018.26.

Windsor, L. C., J. G. Cupit, and A. J. Windsor. 2019. "Automated content analysis across six languages." *PloS One* 14 (11): e0224425. DOI: 10.1371/journal.pone.0224425.