Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 5.2 (2023) 1–34
HTTPS://DOI.ORG/10.5117/CCR2023.2.6.MATE

Machine Translation as an Underrated Ingredient? Solving Classification Tasks with Large Language Models for Comparative Research

Ákos Máté
Centre for Social Sciences, Budapest

Miklós Sebők
Centre for Social Sciences, Budapest

Łukasz Wordliczek
Jagiellonian University, Kraków

Dariusz Stolicki
Jagiellonian University, Kraków

Ádám Feldmann
University of Pécs, Pécs and Óbuda University, Budapest

Abstract
While large language models have revolutionised computational text analysis methods, the field is still tilted towards English language resources. Even as there are pre-trained models for some "smaller" languages, the coverage is far from universal, and pre-training large language models is an expensive and complicated task. This uneven language coverage limits comparative social research in terms of its geographical and linguistic scope. We propose a solution that sidesteps these issues by leveraging transfer learning and open-source machine translation. We use English as a bridge language between Hungarian and Polish bills and laws to solve a classification task related to the Comparative Agendas Project (CAP) coding scheme. Using the Hungarian corpus as training data for model fine-tuning, we categorise the Polish laws into 20 CAP categories. In doing so, we compare the performance of Transformer-based deep learning models (monolinguals, such as BERT, and multilinguals such as XLM-RoBERTa) and machine learning algorithms (e.g., SVM). Results show that the fine-tuned large language models outperform the traditional supervised learning benchmarks but are themselves surpassed by the machine translation approach. Overall, the proposed solution demonstrates a viable option for applying a transfer learning framework for low-resource languages and achieving state-of-the-art results without requiring expensive pre-training.

Keywords: Machine learning, Deep learning, Natural language processing, Classification, Policy topics, Comparative Agendas Project

Introduction

Recent developments in the Natural Language Processing (NLP) field have been nothing short of revolutionary with the widespread adoption of Transformer large language models (LLMs). These neural network-based models took the field by storm, outperforming most of the previously applied supervised methods in a wide range of tasks (Devlin et al., 2018). The text-as-data approach (Grimmer & Stewart, 2013) has now permeated the mainstream of political science, economics, sociology, and communication studies, thus proving to be a truly multidisciplinary methodology (Lind et al., 2021; Theocharis & Jungherr, 2020; Wilkerson & Casas, 2017). Using these tools allowed researchers to process and analyse text data on a large scale, exploiting the deluge of text data and the rapid development of methods and software implementation.

However, like the natural language processing community, social science research is biased towards high-resource languages (e.g., English, other Western European languages, Chinese, and Japanese). In contrast, low-resource languages are more challenging to use in research due to the scarcity of NLP tools available (Baden et al., 2022). Applying machine translation is a widely used approach in the political science literature to solve this resource problem. This approach was demonstrated to be effective for machine learning applications as the machine translated text using Google Translate API comes close to the gold standard of expert translation (De Vries et al., 2018).

In this article, we propose adapting this translation-based workflow to incorporate large language models. As Transformer models are pre-trained in one or more languages, their usage is limited to these domains and texts outside of these languages need to be translated to a bridge language that is part of the pre-training data of the model. While pre-trained models are available for some low-resource languages, the extent of their pre-training may differ, and they might not be available at all for all the languages involved in the comparative research. There are also multilingual pre-trained models available, but their performance (at the time of writing) may trail that of single-language models given the research context. As models are expected to grow in size and complexity, it may be the case that relying on translation is only a stopgap measure until, eventually, sufficiently capable multilingual models are developed with sizeable training data from low-resource languages (Kaplan et al., 2020). However, the timing of these developments is uncertain, and it is important to leverage ready-to-be-deployed tools in comparative research to harness recent advances in natural language processing.

We compare these various types of fine-tuned large language models on a common text classification task in comparative politics: assigning one, and only one, public policy topic (or category, such as education or macroeconomics) to units based on the Comparative Agendas Project (CAP) coding scheme covering 21 such "major topics". The article has a dual aim. First, we intend to assess the potential performance gains using deep learning-based large language models on this typical comparative politics multi-class classification task versus well-established machine learning algorithms. Second, we showcase a possible workflow that leverages machine translation and transfer learning to allow classification across language domains between two low-resource languages (in our use case: Hungarian and Polish).

We assume that this setup (using machine translated input data and the state-of-the-art English language monolingual models) yields competitive classification results with those of small resource language monolingual models (such as an available Polish model, which was pre-trained on limited data) and multilingual models. An additional competitive advantage of the machine translation setup is its relatively low cost and good scalability vis-á-vis pre-training new monolingual or multilingual LLMs. We apply the fine-tuned large language models to the standard CAP text classification task into various topics or categories. While this task is relatively simple compared to question answering or text generation, it is still widely used in social science literature.

The transfer learning aspect of large language models happens during the fine-tuning process, which updates the pre-trained model's general language understanding with the domain-specific context (in our use case: laws and bills). The involvement of machine translation solves the cross-language domain barrier and serves as an (at least) second-best approach whenever monolingual models are unavailable for the given language (or multilingual models are not providing adequate performance). Overall, this approach allows for fine-tuning large pre-trained models, which means lower overall costs and better performance compared to traditional machine learning approaches and (certainly on the cost, but also possibly on the performance and reliability part) human coders.

The proposed workflow demonstrates a viable solution for applying a transfer learning framework for low-resource languages and achieving state-of-the-art classification results with minimal start-up costs. Our results show that the fine-tuned Transformer models outperform the supervised learning benchmarks (Support Vector Machines, Naïve Bayes, and Random Forest) in precision, recall, and accuracy. Moreover, we find that in the case

of low training data sizes (often associated with low-resource languages), the monolingual English BERT model combined with machine translation outperforms the larger multilingual models (such as XLM-RoBERTa) on the benchmark of macro F1 scores. In sum, using English as a bridge language ameliorates the data scarcity problem in low-resource language contexts.

In what follows, we first give a brief overview of the applications of supervised text classification in the literature, showing the path from bag-of-words models to the current neural network architectures. Next, we present the training and target data and model selection. The subsequent section details the results of various models on the target corpus, showing that our workflow using machine translation and large language models is a valid solution to this sort of resource bottleneck in comparative research. The article concludes by assessing the relative merits of the proposed deep learning- and machine translation-based approach vis-á-vis bag-of-words and monolingual models.

Leveraging the Transformer revolution in comparative research

Machine learning has been part of the political science toolbox for over three decades (Schrodt, 1990, 1991). The applications of this toolkit range from methodologically focused works (Montgomery et al., 2012), and conflict studies (Perry, 2013) to electoral and voting studies (Peterson & Spirling, 2018). The text-as-data approach has gained momentum in political science in the past decade (Cardie & Wilkerson, 2008; Grimmer & Stewart, 2013; Monroe & Schrodt, 2008; Wilkerson & Casas, 2017). While a wide variety of research now uses supervised learning as part of the research design, there is a limited selection of studies which use text-as-data in a comparative setting and even fewer using the latest development: Transformer LLMs.

One of the widely used comparative coding schemes in political science is the codebook of the Comparative Agendas Project. In this system, one document (e.g., a law or a newspaper article) is assigned one and only one major policy topic code, such as macroeconomics or health care, out of a total of 21. There have been cases where supervised techniques were used to classify documents using the Comparative Agendas Project codebook, but the majority of such projects still rely on expert coders (a point made by, e.g., Loftis and Mortensen (2020) and Sebők and Kacsuk (2021)). Recent research using supervised learning and the CAP major topic categories use Support Vector Machines, Random Forest, Logistic Regression or Naïve Bayes (Barberá et al., 2021; Dun et al., 2021; Hillard et al., 2008; Loftis & Mortensen,

2020; Sebők & Kacsuk, 2021). As the performance of the bag-of-words models started to plateau, social scientists started to branch out towards more complex representations of texts, such as word embeddings (Rodriguez & Spirling, 2021). In recent research, multilingual sentence embedding was successfully applied to manifesto classification tasks as results surpassed the traditional machine learning models' performance on translated texts (Licht, 2023).

Current state-of-the-art results are achieved using large pre-trained language models. New architectures based on multi-layer bidirectional Transformers (Vaswani et al., 2017) were deployed in speech analysis (Latif et al., 2018; Yang et al., 2017) and various NLP problems (McCann et al., 2018; Pikuliak et al., 2021; Radford et al., 2018; Raffel et al., 2020). Eventually, the flexible BERT model ("Bidirectional Encoder Representations from Transformers") was presented by Devlin et al. (2018), demarcating a turning point in NLP development and opening up the era of foundational models that are "the common basis from which many task-specific models are built via adaptation" (Bommasani et al., 2021).

Transformer architectures leverage the attention mechanism introduced by Vaswani et al. (2017), which allows the model to create weights according to the importance of a given token in its context.[1] While an increasing amount of Transformer models are released on an ongoing basis, our focus here is on the seminal BERT model and its variants, as they have a proven track record when fine-tuned for multi-class classification tasks. By using the bidirectional self-attention mechanism, BERT can utilise each word's context to a greater extent. The bidirectional part means that the model can better understand the role of the word in a given context, leading to improved results.

The BERT model is also available in a multilingual version (often called mBERT) pre-trained on 104 languages. For RoBERTa (Robustly Optimised BERT Pre-training Approach), Liu et al. (2019) re-evaluated the BERT pre-training, and they found that better model performance depends on longer pre-training time, larger batch sizes of data, removing the next sentence prediction phase from the pre-training process, longer sentences and dynamically changing the masking pattern. As a continuation of this work, Conneau et al. (2019) released the XLM-RoBERTa model, a multilingual model that surpassed the mBERT performance on a wide range of benchmarks. It is pre-trained on CommonCrawl data containing 100 languages.

While fine-tuned versions of these models held the record for state-

[1] For an in-depth discussion of the attention mechanism, see Vaswani et al. (2017).

of-the-art results on various tasks at various points, in real-life research contexts, it is not always feasible to fine-tune larger models while, at the same time, smaller models can yield comparable results to their larger variants. The DistilBERT model uses knowledge distillation, compressing a teacher model into a lighter-weight student model by transferring valuable knowledge (Hinton et al., 2015; Kim & Rush, 2016). Distilling has already been used to create smaller and faster versions of large pre-trained models with tolerable loss of efficiency. DistilBERT (Sanh et al., 2019) is a reduced form of BERT with 40% fewer parameters and 97% retained performance.

One major limitation of the language models mentioned above is the limitation on the length of text they can process. With BERT-based models, the upper limit for the input that the model can deal with is 512 tokens, but two tokens are reserved for special tokens (representing the start of a sequence with "[CLS]" and the end with "[SEP]"). The pre-trained BERT-large model has 24 layers (Transformer blocks), a hidden size of 1024, 12 self-attention heads and 340 million total parameters (Devlin et al., 2018).[2] A so-called Longformer model was developed to mitigate this shortcoming and allow for longer input texts (Beltagy et al., 2020).

NLP applications of pre-trained language models also include topic labelling (Béchara et al., 2021). The point of departure for Béchara et al. (2021) was to tackle the outstanding problems of traditional topic models: scalability, human bias in labelling topics, and severely limited replicability. Their suggested remedy implements automatic labelling using already available and recognised codebooks as the model knowledge base to "automatically transfer existing domain-specific knowledge to the process of topic labelling" (Béchara et al., 2021, p. 2). Similar attempts to move forward with transferring knowledge across different languages are few and far between. In one of the existing studies, Pires et al. (2019) show a high level of performance in cross-lingual generalisation. Moreover, what is particularly important for our current research design, their results holds even in the case of a limited topological similarity between given languages.

Another methodological reference point for the current project is the usage of pre-trained language models for knowledge transfer from one corpus to another via fine-tuning. This approach, known simply as transfer learning (Glorot et al., 2011; Thrun, 1998), allows for introducing flexibility in

[2]There is a smaller version called BERT-base, which has 12 layers, hidden size of 768 and 12 self-attention heads, and 110 million total parameters. This model was created mainly to be comparable to the GPT models. However, Devlin et al. (2018, p. 6) reports that "[...] BERT-large significantly outperforms BERT-base across all tasks, especially those with very little training data."

terms of dealing with the feature space of the source and target domains. This framework has been successfully applied in research designs since the mid-1990s (Pan & Yang, 2010; Raffel et al., 2020). For example, Burscher et al. (2015) coded policy issues in news articles and parliamentary questions via supervised machine learning to generalise across contexts. However, their study was based on highly context-dependent training data that decreased the model's generalizability.

Others used Wikipedia document titles as label candidates for topics (Bhatia et al., 2016). Then, they computed word and document embeddings to select the most relevant labels automatically. The approach they named "NETL" (neural embedding topic labelling) outperformed state-of-the-art topic labelling frameworks in terms of simplicity and efficiency. Sun et al. (2019) tried to take advantage of avoiding training a new model from scratch. They applied a three-step fine-tuning solution for BERT for text classification tasks with satisfactory performance. Also, Wu and Dredze (2019) proved BERT to work well when properly fine-tuning its parameters across wide-ranging NLP tasks, from natural language inference, document classification, named entity recognition, and part-of-speech tagging, to dependency parsing.

In this paper, we aim to combine the transfer learning capabilities of the Transformer architecture with neural machine translation (NMT). Machine translation has been shown to work well with machine learning models, and one of our aims is to extend these findings to the newer deep learning models (De Vries et al., 2018). As a new approach, neural machine translation has already shown promising results that outperform the previous solutions of Statistical Machine Translation (Bojar et al., 2016). From a probabilistic perspective, translation is equivalent to finding a target sentence Y that maximises the conditional probability of Y given a source sentence X. NMT approaches fit parameterised models to maximise the conditional probability of sentence pairs using a parallel training corpus.

Once the conditional distribution is learned by a translation model, given a source sentence, a corresponding translation can be generated by searching for the sentence that maximises the conditional probability. As Sutskever et al. (2014) have found, recurrent neural network (RNN) based machine translation models using LSTM (long short-term memory) units can achieve similar results to phrase-based translation systems. Cho et al. (2014) presented an encoder-decoder-based solution which further improved the performance of neural machine translators. Bahdanau et al. (2014) used an attention mechanism instead of RRN layers, further improving the effi-

ciency of NMT. Building on the recent seminal contributions by Vaswani et al. (2017), transformer-based solutions have become the most widely used NMT models, using elements of both the encoder-decoder architecture and the attention mechanism to outperform previous solutions.

As this brief overview shows, the intersection of the NLP and social science field has been using various text-as-data methods extensively; however, the application of pre-trained language models has been a relative rarity in this literature. We aim to contribute to narrowing this gap by showing a possible way to leverage new methodological developments and providing a research design to apply these models in a comparative setting.

Data and methods

Fine-tuning data

In crafting our research design, our goal was to create a classification pipeline that allows us to leverage the transfer learning strengths of the pre-trained and fine-tuned large language models to cover the language barrier (via machine translation) and solve the underlying CAP-style policy topic classification task. We evaluate four methodological approaches for this task: traditional machine learning algorithms, input language monolingual LLMs (such as a Polish BERT), multilingual models and our proposed setup of combining machine translated input texts and state-of-the-art English language monolingual models.

For testing purposes, we used Hungarian laws and bills (introduced but not necessarily adopted legislative texts) as the training dataset for fine-tuning the Transformer models (bar PolBERT, see below). Based on prior research (Sebők et al., 2020), we assumed that each bill has the same label as the enacted law, so the bills inherited their labels without additional need for human coding (this means that, on average, even though the text of the bill changed during the legislative process, its overall policy emphasis did not). As for the training corpus for fine-tuning the language models, we used a corpus containing a total of 7794 units (4088 laws and 3706 bills).[3] The documents' timeframe was between 1990 and 2018.

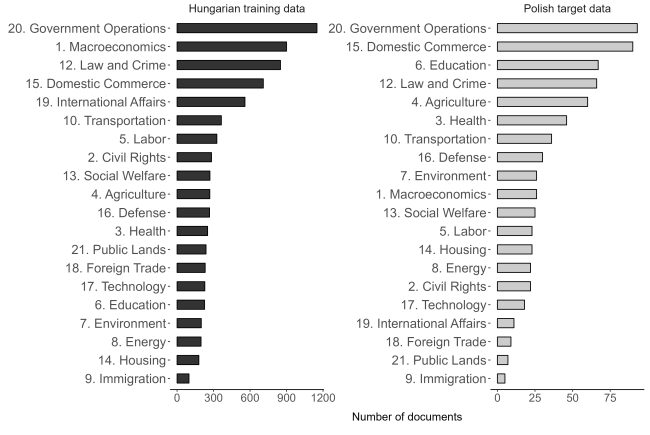
There are 20 Comparative Agendas Project major topic labels in our dataset ranging from education to defense.[4] The distribution of categories in the hand-coded Hungarian data is highly uneven, as not every domain

[3]The labelling of the data was done on the original documents by native Hungarian speakers, with a minimum of two coders for each item.

[4]The dataset available did not contain the category of "Culture" (21).

received equal attention from lawmakers. It is well-documented in the literature that this issue persists across domains as CAP-coded media corpora are naturally imbalanced (Boydstun, 2013; Sebők & Kacsuk, 2021). The distribution of documents across major topics (both for the training and target data) is shown on Figure 1.

Figure 1: Major topic distribution in hand-coded training and target data



We used this hand-coded dataset to fine-tune the Transformer models. Pre-processing steps included cutting the boilerplate sections (using regular expressions) from the beginning of the documents. Eliminating the boilerplate is necessary to maximise the information content of the first 512 tokens. The BERT tokenizer algorithm splits words into multiple tokens (including punctuation).[5] We applied the standard pre-processing steps for the machine learning models competing with LLMs: lowercasing, eliminating punctuation, special characters, numbers, and stop words (Grimmer et al., 2022). We also used TF-IDF weighting when creating the document-feature matrix from the corpus.

Target data

The Polish target dataset consists of all statutes in force as of September 1, 2021, except those consisting only of amendments to other laws and those

[5]For example, the BERT tokenizer from the Transformers library will tokenize the "GPU" word into ["gp", "##u"] tokens, as GPU is not part of the BERT tokenizer's vocabulary. The "##" indicates that the token is split from the preceding one.

giving parliamentary consent to ratification of international agreements (due to their boilerplate nature). While the selection criteria for the training and target data differ (date of adoption vs date in force), that should have no real effect on the meaningfulness of our experiment. The key is applying the same coding system to similar texts (laws and bills) in comparable political systems.

The statutes and their official consolidated texts have been scraped from the ISAP database (the Internet System of Legal Acts), with texts processed using the any2txt python package. The starting target dataset consisted of 1014 items and covered the totality of parliamentary legislation in effect in September 2021. In line with the literature, we consider the gold standard for such classification exercises to be double-blind expert coding. Therefore, we selected the 705 pieces of legislation where such an agreement had been reached between at least two independently working expert coders.

It is also notable that individual proposals' policy coherence (as well as their length) may vary drastically. This means that a certain level of disagreement over the dominant policy topic (as there can only be one label per unit) of the text is natural (as many laws cover a mixture of topics). We mitigated this issue by only considering the laws for which at least two coders agreed. As Figure 1 shows, the target dataset's distribution is very similar to the training set's: it is highly imbalanced between categories. There are some discrepancies between the training and target set in terms of category sizes, but the largest category for both is Government Operations and the smallest is Immigration. As such, the datasets pass a face validity test. Furthermore, the descriptive statistics of the two datasets are also reasonably resembling, with units of the Polish target data being slightly longer on average (see Table 1).

Table 1: Descriptive statistics of the training and target datasets

	Sample size	**Median word count**	**Std.dev. word count**
Hungarian training data	7794	585	204.92
Polish target data	705	612	174.76

Model selection and functions

The article's main goal is to evaluate four available options to solve the classification task related to a language with limited NLP resources (Polish) when assigning a single policy topic to each law and bill in our target dataset. Tra-

ditional machine learning algorithms, input language monolingual LLMs (such as a Polish BERT), multilingual models and our proposed setup of combining machine translated input texts and state-of-the-art English language monolingual models were deployed to establish a performance rank order amongst them.

In the first step, to limit the number of LLMs deployed, we ran experiments on the bigger hand-coded Hungarian source dataset used for fine-tuning most of the models. Table 2 provides an overview of all models used in this article and their respective functions in the research design (whether they serve a benchmarking or evaluation role). The benchmarking phase (see first seven models) yielded valuable insights into which models to keep for the later evaluation steps. The choice between various models partly hinges on a trade-off between performance and available computing resources. The BERT model is pre-trained on a large corpus of the BooksCorpus, with 800 million words (Zhu et al., 2015) and English Wikipedia, with 2500 million words. We accessed the pre-trained model through Hugging Face's Transformers library (Wolf et al., 2019).

Table 2: The large language models used and their functions

Model	Size	Language	Fine-tune data	Target
			Benchmarking	
BERT	Base	EN	HU translated to EN	HU translated to EN
BERT	Large	EN	HU translated to EN	HU translated to EN
HuBERT	Base	HU	HU	HU
RoBERTa	Base	EN	HU translated to EN	HU translated to EN
DistilBERT	Base	EN	HU translated to EN	HU translated to EN
XLM-RoBERTa	Large	Multi (100)	HU, HU translated to EN	HU, HU translated to EN
			Evaluation: Monolingual vs. Multilingual	
PolBERT	Base	PL	PL original	PL original
XLM-RoBERTa	Large	Multi (100)	PL, HU original	PL, HU original
mBERT	Base	Multi (104)	PL, HU original	PL, HU original
			Evaluation: Original language vs Machine translated	
BERT	Base	EN	HU translated to EN	PL translated to EN
BERT	Large	EN	HU translated to EN	PL translated to EN
HuBERT	Base	HU	HU	PL translated to HU

Using a pre-trained model means that the heavy lifting of the unsupervised pre-training using a large amount of computational resources and training data is already done, and we can rely on this model to add our (signif-

icantly smaller) domain specific corpus to fine-tune the model. Fine-tuning relies on the same architecture and is considerably faster and cheaper than pre-training–this constitutes the first added value of deep learning models that we leverage in our research design. Table 2, therefore, also provides information on the data used for fine-tuning the model as, on multiple occasions, the same model was fine-tuned on different data sources.

In order to establish benchmarks for the evaluation phases of our research design (on the Polish target data), we first compared the performance of the large and base variants of the English-language monolingual BERT, huBERT (a comparable model pre-trained on Hungarian data), as well as the RoBERTa and DistilBERT models strictly on the Hungarian language input dataset. The multilingual XLM-RoBERTa (in the large version) was also used. These benchmarks helped evaluate model performance for the baseline (input language monolingual, multilingual) and the machine translated track of our research design.

Based on the results of the benchmarking phase (see below), we selected a more limited set of models for the evaluation phases. For the first evaluation phase related to comparing monolingual and multilingual model performance on the Polish target data, we used a (base-sized) Polish BERT model (PolBERT – Kłeczek (2020)),[6] as well as some widely used multilingual models, including XLM-RoBERTa (large) and mBERT (only base version exists). The Polish and Hungarian BERT models match the BERT-base model in their architectures (containing 110 million parameters); however, they use considerably smaller training data. The original BERT-base model was pre-trained on a 3,3 billion word corpus, whereas the Polish model used 1,8 billion, and the Hungarian model around 1 billion words.[7]

Bridging the cross-language domain with machine translation

In the second phase of the evaluation, we also added English language models to the competition after the supplementary step of machine translation. Our workflow is charted in Figure 2. The key problem that this combination of language models and machine translation addresses is that in comparative research designs for low-resource languages, there may not be a pre-

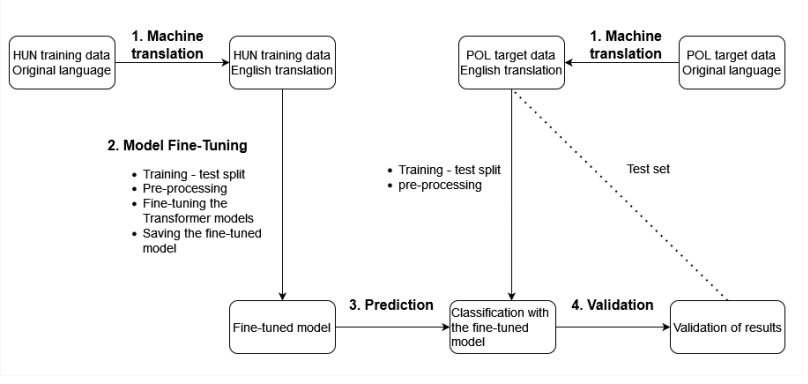
[6]The Polish model was pre-trained on the Polish subset of ParaCrawl, Open Subtitles, Polish Wikipedia snapshot, and the Polish Parliamentary Corpus. More details can be found in its HuggingFace repository: https://huggingface.co/dkleczek/bert-base-polish-cased-v1..

[7]For details on the original BERT pretraining data see Devlin et al. (2018), for HuBERT's see Nemeskey (2020).

trained LLM available or available models might lack scale regarding their pre-training. We used the English language as a common denominator between the different source languages to exploit the strengths of Transformer models. De Vries et al. (2018) demonstrated that using machine translation is a valid tool for comparative researchers as using translated text for machine learning models yields similar results to gold standard human translation. Based on our results, this insight also carries over to research designs based on modern large language models.

In Step 1, the open-source neural machine translation (NMT) model called Opus-MT was used to translate the Hungarian and Polish documents to English (Tiedemann & Thottingal, 2020). We decided on Opus-MT as an NMT model since it is an open-source, non-commercial project that is freely available regardless of the document length that needs to be translated. Based on our measurements, it achieves similar results to popular translation models, such as Google Translate and DeepL, without the costs of scaled-up projects (see Table B1 in the Appendix).

Figure 2: The classification workflow using machine translation and Transformer models



In Step 2, during the fine-tuning step, we tokenised our documents and limited the length of each document to 512 tokens.[8] Contrary to machine learning algorithms, which use document-feature matrices as inputs, Transformer models do not require extensive pre-processing as the model utilises all of the contextual environment of each token. Fine-tuning the BERT model uses the same architecture as the pre-training step with the important distinction that pre-training is unsupervised, and fine-tuning is

[8]The text pre-processing and tokenization was done with the transformers library (Wolf et al., 2019).

a supervised process. In Step 3, we saved our fine-tuned model and used it to classify the translated Polish laws. Finally, in Step 4, we evaluated the out-of-sample performance of our models using the test dataset set aside from our original translated corpus of Polish laws.

Benchmark results for the source data and model selection

In order to establish expectations as to how well various large language models would perform compared to frequently used supervised learning methods in the literature, we ran a series of tests on the comparatively larger dataset for Hungary. This exercise allowed for the narrowing down of the modelling options for the virgin Polish data based on empirical results for another "small" language with relatively limited language technology resources.

First, we carried out the classification exercise using various bag-of-word models. Since these models do not limit the size of the inputs, we used the full documents. These baseline models included the Naïve Bayes classification (NB), Support Vector Machines (SVM), and Random Forests. In all three cases, we applied the customary bag-of-word pre-processing steps: lowercasing, removing stop words, numbers, punctuation, line breaks and other superfluous whitespaces. In each case, we applied TF-IDF weighting to the document-feature matrix.[9] In order to focus on LLMs, we limited the application of machine learning algorithms to the translated texts to establish benchmarks for the relative performance of the translation-based workflow of Figure 2 (we expect that these algorithms will have suboptimal results vis-á-vis LLMs regardless of input language).

In order to decide which Transformer models to use on the target language (Polish) data, we also measured their fine-tuned performance on the source dataset (original and English-translated Hungarian bills and laws). We fine-tuned and validated 6 models (BERT Large, BERT Base, HuBERT, RoBERTa, XLM-RoBERTa Large and DistilBERT).[10] Out of these, HuBERT and the multilingual XLM-RoBERTa models were tested on original Hungarian data, while all models (except for HuBERT) were also run for the English-translated Hungarian dataset. Adding HuBERT (Nemeskey, 2020) to the mix allowed for directly gauging the information lost in translation

[9]The application of the machine learning models and the TF-IDF weighting was handled with the scikit-learn library (Pedregosa et al., 2011).

[10]The BERT-based models were fine-tuned on the first 512 tokens of the documents. We also carried out robustness checks for various input data (see below – detailed results are presented in Appendix A). We used uncased versions for all models.

and learning how well smaller pre-trained BERT models for low-resource languages perform compared to the larger English BERT models.[11]

It is also important to note that several hyperparameters affect model performance. We found that the most impactful parameters are batch size (the amount of data used for a round of input to the neural network) and the learning rate for the Adam optimiser with decoupled weight decay. The role of the optimiser is to update the neural network's weights during the training process. In line with findings on fine-tuning BERT models, we randomised the training data order and experimented with various random seeds as part of the hyperparameter tuning process (Dodge et al., 2020). The final pre-trained models were initialised with a learning rate of 2e-5, batch size of 4, and a dropout rate of 0.1 for the final classifier layer to avoid overfitting. We fine-tuned the models for 3 epochs using a Tesla V100 GPU (available hardware resources limited the parameter-tuning experiments).

Table 3 compares the performance of machine learning models and deep learning-based LLMs for the classification task on (original or English-translated) Hungarian source data. It shows the dominance of the Transformer models compared to traditional machine learning approaches. Depending on evaluation metrics, there are some exceptions, such as the F1 score provided by SVM and the precision achieved by the Random Forest. However, SVM only surpasses the multilingual XLM model applied to translated data, a suboptimal combination. Furthermore, in the case of high Random Forest precision, this result is achieved in a trade-off with recall, resulting in the second-lowest F1 score.

Overall, the machine learning models' performance is similar to studies using similar data and coding schemes. Recent research using the CAP major topic categories achieved a precision between 0.65 and 0.71, using Random Forest, Logistic Regression or Naïve Bayes (Barberá et al., 2021; Dun et al., 2021; Loftis & Mortensen, 2020). Further evidence shows that dedicating considerable computational resources to single language classification using a voting ensemble of SVM models plateaus around a precision of 0.85 and a recall of 0.60 (Sebők & Kacsuk, 2021; Sebők et al., 2021).

When it comes to the selection of LLMs, we implemented widely used models with both original and translated data (where applicable). Regarding F1, HuBERT performed best, with XLM-RoBERTa on the original language data coming in a close second. RoBERTa and the two BERTs showed similar results. The remaining deep learning and machine learning models were

[11]The Hungarian model was pre-trained on the Hungarian subset of the Common Crawl data (including a snapshot of the Hungarian Wikipedia). For more details, see Nemeskey (2020).

Table 3: Benchmarking: Macro F1 performance of models on Hungarian data

Model	Accuracy	Precision	Recall	Macro F1
HuBERT-base HU → HU	0.825	0.831	0.826	0.825
XLM-RoBERTa HU → HU	0.840	0.830	0.820	0.819
RoBERTa hu-EN → hu-EN	0.819	0.823	0.820	0.816
BERT-large hu-EN → hu-EN	0.818	0.811	0.825	0.814
BERT-base hu-EN → hu-EN	0.806	0.814	0.786	0.787
DistilBERT hu-EN → hu-EN	0.799	0.782	0.780	0.778
Support Vector Machines hu-EN → hu-EN	0.753	0.743	0.758	0.747
XLM-RoBERTa hu-EN → hu-EN	0.770	0.740	0.740	0.736
Random Forest hu-EN → hu-EN	0.714	0.853	0.638	0.700
Naïve Bayes hu-EN → hu-EN	0.664	0.703	0.666	0.652

Note: The model name indicates the transformer model, its size, translation direction in the training data and translation direction in the target or test data. E.g., "BERT-large hu-EN → hu-EN" is a BERT transformer, large size, using English training data translated from Hungarian (lowercase indicates the provenance of the data, and uppercase indicates the language that it was translated into), used to predict labels for test data translated from Hungarian to English.

relegated to the lower half of the table (DistilBERT and XLM-RoBERTa on translated data). The models in Table 3 in bold were selected for evaluation on the Polish target data. We retained all machine learning models to gain additional insights into their suboptimal performance, and we also kept the best-performing LLMs (except for RoBERTa, which showed almost identical performance as the BERT large).

Evaluation results for the target data

Model comparison

We conducted a multi-step experiment to assess the performance of the four competing approaches on the low-resource Polish dataset (with significantly more limited fine-tuning data than in the Hungarian case). In the first step, we compared six Transformer models and three machine learning models (see Table 4). The results show that Transformer models substantially outperform the popular machine learning approaches. This is mainly a function of the difficulty of the task at hand. A major drawback of the bag-of-words-based models is that all contextual information is lost, and if the target data contains a significantly different set of words than the training

data, poor performance is likely.

Table 4: Performance on the Polish target data

Model	Accuracy	Precision	Recall	Macro F1
BERT-large hu-EN → pl-EN	0.777	0.747	0.766	0.750
BERT-base hu-EN → pl-EN	0.780	0.714	0.762	0.727
HuBERT-base HU → pl-HU	0.780	0.570	0.580	0.559
mBERT-base PL → PL	0.610	0.571	0.567	0.554
Support Vector Machines hu-EN → pl-EN	0.552	0.563	0.500	0.503
Naïve Bayes hu-EN → pl-EN	0.587	0.581	0.539	0.501
PolBERT-base PL → PL	0.704	0.489	0.486	0.475
XLM-RoBERTa-large PL → PL	0.610	0.350	0.470	0.379
Random Forest hu-EN → pl-EN	0.365	0.537	0.229	0.258

Note: The model name indicates the transformer model, its size, translation direction in the training data and translation direction in the target or test data. E.g., "HuBERT HU → pl-HU" is a HuBERT transformer, base size, using original Hungarian training data (no translation) used to predict labels for test data translated from Polish to Hungarian (lowercase indicates the provenance of the data, and uppercase indicates the language that it was translated into).

There are also significant differences between the Transformer results. Using the BERT-base models, we can directly compare the performance of a multilingual model, an English and an original (small resource) language model (PolBERT). The results show that using the English BERT model on the translated data yields the best results with a 0.75 macro F1 score. Significantly, all other LLMs performed drastically worse, with HuBERT in third place with a gap of 0.168 in macro F1.

To apply the HuBERT model, we translated the Polish data into Hungarian. This creates an additional choice which lowers translation costs compared to the bridge language option (as only one set of translations is necessary). Notably, the model's performance is markedly lower than the one tested with the original Hungarian data. This drop in the macro F1 score can result from the poorer machine translation quality between Polish and Hungarian.[12] While multilingual models (tested on the original Polish data–mBERT and XLM-RoBERTa) also outperform the machine learning alternatives, they still lag in performance versus the machine translation-based alternative setup. Contrary to our expectations (based on the Hun-

[12]For translating the Polish laws to Hungarian, we used the Google Translate API, as the Opus-MT model family does not have a Hungarian-Polish model trained. See also Appendix B.

garian data results in Table 3), the XLM-RoBERTa model underperformed the multilingual BERT model.

An additional notable result is that, in contrast to the Hungarian BERT model's performance, the Polish BERT model significantly underperformed the fine-tuned English model (despite being pre-trained on a Polish parliamentary corpus, among other sources). Furthermore, possibly due to the dearth of training data for certain individual classes, the class-based predictions behind the average F1 score contain 8 zeros with the obvious impact of decreasing the average level (see more on this below). Using a weighted average for the F1 score of PolBERT (which accounts for the small test sample size) yields an F1 score of 0.67. In general, these smaller pre-trained models (vis-á-vis even the English language BERT base) are more heavily impacted by the lack of extensive fine-tuning data (as indicated by the difference between the size of the Hungarian and Polish training data).

Overall, a key takeaway is that the performance of various pre-trained Transformer models varies significantly. This may be attributed to a slew of factors. Larger models tend to outperform their smaller counterparts (as in the case of BERT large and base). The superiority of models applied to translated data may also be due to the higher relative quality of machine translation to the quality of pre-training (which may be limited for small resource languages for both monolingual and multilingual models). Furthermore, the main bottleneck for all original language classification tasks appears to be the small sample size for labelled data, for which even larger multilingual models cannot compensate. We return to these issues in the Discussion.

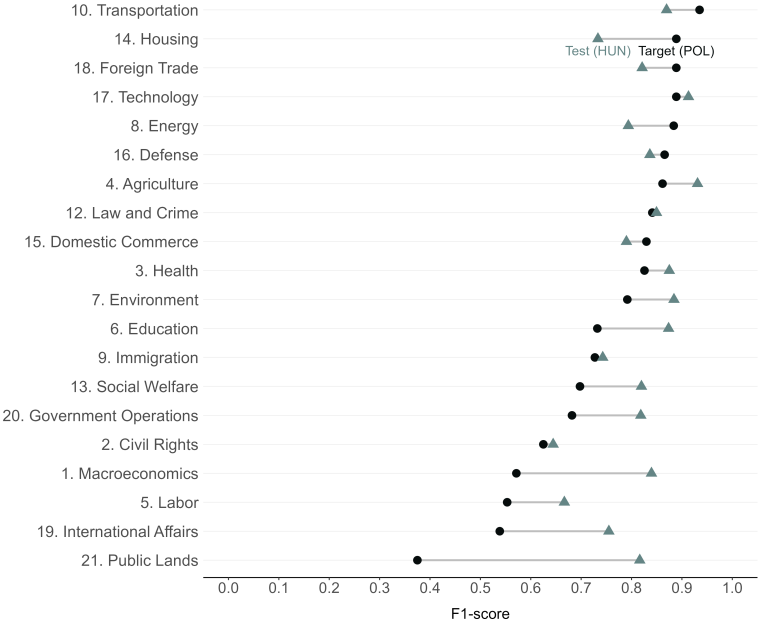
Class-level evaluation

With class-level performance analysis, additional insights into the factors affecting model performance may be gained. Examining the results of the state-of-the-art fine-tuned BERT Large model (fine-tuned with the first 512 tokens), Figure 3 shows that for 6 categories out of the 20, the F1 scores of the model are better on the Polish target documents than on the Hungarian out-of-sample test set.[13] Notably, only 5 categories are below an F1 score of 0.7, and there are 11 categories with a score above 0.8. On the one hand, these results show that while, on average, there is some performance loss when switching domains (from Hungarian laws and bills to Polish laws), this is not equally distributed between the categories.

[13]These are: "Transportation", "Housing", "Foreign Trade", "Energy", "Defense", and "Domestic Commerce".

On the other hand, the "Public Lands", "International Affairs", and "Macroeconomics" categories have the poorest F1 scores in the target data while having high F1 scores in the test data. The large discrepancy in these categories might result from varying policy emphases and political idiosyncrasies. Model performance, in this case, is undoubtedly a function of available data, as categories with the most documents have better F1 scores.[14]

Figure 3: Class-level performance differences for target and test data



Robustness checks for input data

We also experimented with various sampling methods and extractive summary approaches to see how data input impacts model performance. A handy option for augmenting a small input dataset is to leverage more features by adopting a so-called "sliding window" technique that rolls in the text in batches. Alternatively, text summarisation algorithms can be utilised for condensing longer texts. In order to provide robustness checks for the

[14]We checked this relationship with a bivariate regression and sample size had a statistically significant effect on the F1 scores.

above results, we experimented with these alternative input provision techniques. As in most similar research, the baseline here is a truncated dataset containing only 512 tokens from each document.

As a first alternative, based on the sampling test results, we created a five to ten-sentence-long summary of each document in the training data and used it as input. The extractive summaries of the original documents were composed using the LexRank algorithm, a stochastic graph-based method that relies on Eigen-vector centrality in a network of sentences (Erkan & Radev, 2004). As a second alternative, to expose the whole training data to the model, we sliced the documents into 512 token length subsets (resulting in 229 390 observations). This implements the 'sliding window' solution, as the fine-tuning process allows the model to ingest the whole document in 512 token increments (Ding et al., 2020).

Table 5 shows the results for the Polish input data translated to English and fine-tuned for the classification task with an English language BERT large model. Results show that the baseline approach of using the first 512 tokens yields the best performance, with extractive summaries coming in second and the sliding window method a distant third. One explanation for the worse results for this latter may be that using the sliding window training data may result in overfitting the model (despite using dropout and randomised input during the fine-tuning). Additionally, it looks likely that the starting segments of bills and laws provide a good prediction of the main policy emphases of the entire document. These results also suggest that creating a summary version of the laws dilutes the information content and/or adds noise to the analysis and should be considered a secondary choice vis-á-vis the 512 tokens approach (or, for specific use cases, even the sliding window method).

Table 5: Input data-based robustness checks for BERT large

Input data	Accuracy	Precision	Recall	F1
512 tokens	0.777	0.747	0.766	0.750
Extractive summary	0.757	0.718	0.734	0.714
Sliding window	0.677	0.675	0.630	0.622

In order to better tease out the language- and training-set size-specific aspects of these results, we also ran additional experiments on the Hungarian data to evaluate the robustness of our results (see Appendix A and B). In Appendix A, we supplemented the robustness checks regarding how various training data forms (512 tokens, extractive summary, sliding window)

affect various LLMs for the Polish dataset with an analysis of the Hungarian dataset. The results in Table A1 show that for the translated Hungarian data, the difference in F1 scores is relatively minor, with the sliding window approach yielding the best macro F1 scores (except for RoBERTa). The 512 tokens approach comes in second for all models except for RoBERTa, which performs best with this input. The most significant gap is for the original Hungarian data, where the sliding window approach yields a 0.9 level performance instead of the 512 token input's 0.83. Depending on concrete input data and research designs, settling on either the 512 tokens or the sliding window approach should provide a safe choice.

As an additional robustness check, in Appendix B, we show results with 4 different versions of fine-tuning data that differ in category balances. The tests show that data imbalance in the fine-tuning training set is not a severe impediment to model performance. This also indicates that there is no need for additional data augmentation or down-sampling in this case, as the results are not significantly affected.

In the final robustness check for our machine translation-based workflow, we benchmarked the performance of the Opus ML model to two widely used alternatives, Google Translate and DeepL (see Appendix B). Using all three services, we translated 98 Hungarian documents to English and found that Opus's mean similarity values (as measured in cosine similarity) were close to those of the widely used commercial translation services. We realised this similarity in output while leveraging the advantages of an open-source project which can be deployed at no cost. These results further buttress the viability of the proposed workflow for low-resource languages and comparative projects with limited funding.

Discussion

The goal of this article was to demonstrate how the latest developments in the field of natural language processing can be leveraged in comparative social research. Recently developed Transformer or large language models, with the proper fine-tuning for the specific task at hand, allow for lower overall cost and better performance than traditional machine learning approaches. To demonstrate the added value of these models in real-life research processes, we used several language models to apply transfer learning between translated Hungarian and Polish law texts using the Comparative Agendas Project coding scheme. The language models were used to classify legal documents (bills or laws) into one of the 20 CAP categories. While, this task is not the most computationally complex or expensive (e.g., compared

to question answering) in the NLP toolbox, it is nonetheless an important step for comparative political research.

Our results showed that deploying such LLMs in various forms, such as (a) multilingual models or monolingual models via (b) either a bridge language or (c) directly (in case they are available and to the extent of their training), served as a viable alternative to both hand-coding and traditional machine learning algorithms. Given the similarity of these three options on a Hungarian project used for benchmarking, we used a similar low-resource language (Polish) for a shoot-out test to gauge their applicability in a comparative research design.

The results for the Polish project showed that using the English BERT model on the translated data yielded the best results in terms of macro F1 score – and by a wide margin. This seemingly counterintuitive result points to the current limitations of both monolingual and multilingual models for so-called low-resource languages. Ranathunga et al. (2023) define low-resource languages as "under resourced, low density, resource poor, low data, or less resourced". Joshi et al. (2020) classify languages into 6 categories, from high-resource to low-resource. Here, the low-resource categories are characterised by small or no labelled datasets, some unlabelled data (but labelled data collection is challenging) and with some language support community present. Examples of these categories include Slovene, Irish, and Nepali. Based on this categorisation, Polish and Hungarian can be well classified as low-resource languages.

However, the two projects in our research design show a different degree of this resource limitation. The Hungarian benchmark project relied on a better pre-trained model (trained on more extensive and more varied data based on available descriptions), and the labelled data used for fine-tuning was also significantly larger than in the Polish case. Based on our results, the loss of information due to translation was more than compensated by the gained accuracy of the leveraged (and significantly larger) Hungarian labelled dataset. This points to machine translation as a critical resource for comparative projects on low-resource languages with adequate labelled data for at least a single language/country case of the sample.

Based on our experiment, the main bottleneck for solving the classification task with original language input data is the small labelled data size, for which even larger multilingual models (showing promising results on the Hungarian benchmark case) cannot compensate. As the Hungarian example shows, a relatively large amount of high-quality labelled data can improve performance, even though Hungarian is still a low-resource lan-

guage (in terms of, for instance, the pre-training scope of the HuBERT model used and the limited variety of available models).

We can generalise some findings regarding the advantages and drawbacks of the various modelling approaches based on the empirical results from our experiment. As Table 6 presents, traditional machine learning algorithms, still the default option for many automated classification projects, cannot provide state-of-the-art results on most metrics of interest in the age of LLMs. They have lower entry costs in terms of the learning curve and computing needs,[15] but for projects aimed at producing cutting-edge research, they can no longer serve as the go-to solution for a variety of reasons which contribute to worse performance than other solutions (such as their limited capabilities to handle imbalanced input data).

Regarding the second option of monolingual Transformer models, the critical distinction is between low- and high-resource languages. For low-resource languages, both the pre-training scope of LLMs and the limited availability of labelled data for any given task can prevent projects from achieving state-of-the-art results (see the performance of PolBERT). While they enjoy advantages vis-á-vis multilingual models (state-of-the-art performance given sufficiently large, labelled datasets for fine-tuning), their availability is not universal for all languages needed in comparative projects, especially when it comes to their uniformity (for some languages, a pre-trained large BERT may be available, for others, only a base version).

As for the costs and learning curve associated with this option, the adequate pre-training of monolingual Transformer models is costly and complex as it requires an order of magnitude larger hardware and training data than just the fine-tuning process alone. In contrast, the only step necessary in workflows based on traditional machine learning algorithms is the supervised training of the model using the labelled dataset. Moreover, the fewer resources a language has (in terms of the scope of pre-training and variety of pre-trained models), the less likely it is to provide state-of-the-art results on similar labelled datasets. Furthermore, while machine learning models benefit from accessibility, as they are implemented in most statistical software libraries and their training does not require specialised hardware, the same is less true regarding LLMs.

Our results were also promising when it comes to the third option of multilingual LLMs. For the benchmark Hungarian project, at least one such model (XLM-RoBERTa) produced competitive results with the state-of-the-

[15] Having said that, at the time of writing, many options are available for utilizing free GPU resources for fine-tuning LLMs. One of the notable such free GPU resources is Google's Colab service.

Table 6: Advantages and drawbacks of classification approaches

	Traditional machine learning models	**Monolingual LLMs**	**Multilingual LLMs**	**Machine translation + monolingual English LLM**
Performance for high resource languages	Lags behind the state-of-the-art when used with high dimensionality data (such as texts)	State-of-the-art results	Comparable performance to monolingual LLMs	State-of-the-art performance provided pre-training and fine-tuning data is ample
Performance for low resource languages	Lags behind the state-of-the-art when used with high dimensionality data (such as texts)	Varying performance due to low resources and limited availability	Varying performance due to low resources and limited availability	State-of-the-art performance
Learning curve	More accessible for the social research community	Fine-tuning requires additional training	Fine-tuning requires additional training	Fine-tuning requires additional training
Computing cost	Cheap training (no special hardware need)	Special hardware needs (GPU for fine-tuning), costly pre-training, relatively cheap (one time) fine-tuning	Special hardware needs (GPU for fine-tuning), costly pre-training, relatively cheap (one time) fine-tuning	Special hardware needs (GPU for fine-tuning), costly pre-training, relatively cheap (one time) fine-tuning
Translation cost	None	None	None	May range from free to large cost (but lower cost and better quality is expected over time)
Comparative usage	Scales well: adaptable to any comparative research setup	Language coverage is still scarce	Scales well: adaptable to any comparative research setup where languages are covered by the LLM	Scales well: adaptable to any comparative research setup
Input data sensitivity	More sensitive to data issues (e.g.: unbalanced training data)	Robust to various input data issues	Robust to various input data issues	Robust to various input data issues

art. At the same time, performance dropped using the original Polish data: all tested multilingual models performed in the 0.55 macro F1 score range, significantly below the top performer monolingual models using translated input (0.72-0.75 macro F1). This indicates that multilingual models can perform on par with monolingual models only given high quality and ample labelled data. These findings underscore our contention that the machine translation approach may only be a stopgap measure given the pace of development of LLMs. However, based at least on our current measurements, the universal applicability of multilingual LLMs is still not a reality, especially when it comes to low-resource languages.

These assessments of the three alternatives to the machine translation track already highlight the potential of combining deep learning-based machine translation with LLMs following a similar Transformer design. This approach yielded the best performance on our target Polish dataset for a low-resource language with limited labelled data. One advantage of the translation-based workflow and using English as a bridge language is that this solution scales well for any pair of languages, given that translation is possible. This is a significantly larger scope of languages than what is currently available in the pre-trained LLM offering for low-resource languages.[16]

Moreover, there is no need for the expensive and complex pre-training phase (as the case would be if one were to pre-train a model for a low-resource language) as many cutting-edge English models are publicly available and ready for fine-tuning with domain specific texts. This is good news for comparative researchers who have fairly large, labelled datasets in a low-resource language (such as Hungarian in our example) and want to carry out a comparative analysis using similarly low-resource languages (Polish in our example).

The key drawback of this workflow is that machine translation can be costly for large corpora in terms of monetary cost (if one uses a paid API service such as Google Translate) as well as in terms of computational time. Fine-tuning the models also requires specialised hardware that might be hard to access (but that is no comparative disadvantage vs the other LLM-based approaches). Moreover, given the availability of open-source solutions (such as the one used in this article), this additional cost may be

[16]As an example, Opus-MT supports 187 languages (https://github.com/UKPLab/EasyNMT#Opus-MT), while Google's NMT engine supports 106 language pairs (https://cloud.google.com/translate/docs/languages). The smallest range in our sample of services is offered by DeepL, which supports just 29 languages (https://support.deepl.com/hc/en-us/articles/360019925219-Languages-included-in-DeepL-Pro).

mitigated in concrete research projects depending on a number of factors. The translation step may also lead to a loss of context, but this concern did not bear out in our experiments. Finally, we expect the unit cost of machine translation to decrease over time and quality to increase. These potential factors are reasons to keep the machine translation setup in the toolkit for comparative projects.

In concluding our discussion, it is important to reflect on the dynamic nature of the NLP field. Based on available research corroborated by our findings, Transformer models are a meaningful improvement over the (currently) more widely used bag-of-words-based supervised learning methods (especially in the light of how well they perform with a relatively small training sample). While the underlying Transformers and deep neural networks might be more complex in theory than a regularised regression, implementing these large language models is accessible to the wider research community without the need to implement or pre-train the models from scratch. This pre-trained nature of BERT and its peers' most significant advantage is that fine-tuning can be done on freely available computational resources.

The comparison of the three LLM-based research designs produces less clear-cut judgements. Their practical performance may depend on a variety of factors ranging from the general state of the NLP-community for the target languages to the availability of labelled data. Especially when it comes to multilingual models, their pre-training is more resource intensive as it has to include materials from several languages. However, this performance gap is likely to disappear, and multilingual models may become the norm over time. This would be a welcome development for comparative research. Nevertheless, the critical bottleneck of an uneven availability of labelled fine-tuning data for all target languages may still remain and will warrant the continued search for alternatives.

Our proposed alternative relies on machine translation with all of its benefits and limitations (the circumvention of which is a promising avenue of research in and of itself). Despite these and similar potential deficiencies – and notwithstanding the fact that the computational social science literature is still in the early stages of exploring Transformer based models – our results suggest major untapped potential in applying transfer learning using fine-tuned large language models in comparative politics and beyond. This case may be even stronger for low-resource languages with limited NLP resources in general and task-specific labelled data in particular.

Finally, while our results prove the new approach of Transformer models to be an unmitigated success for a fairly common classification task (the

classification of the policy topic of documents), this new approach is by no means a silver bullet for all comparative classification projects. Usual caveats about topic drift and domain shift still apply even for the Comparative Agendas Project case, let alone for thoroughly different research endeavours. For the case at hand, transfer learning across vastly different parliamentary or legal systems is likely to have poorer results than transfer learning between similar systems, such as the Hungarian and Polish domains. The exploration of these domain boundaries offers a fruitful avenue for future research.

Replication materials: The results can be replicated using the following repository: https://doi.org/10.6084/m9.figshare.24025845.v1

Acknowledgements: The research project was supported by the Ministry of Innovation and Technology NRDI Office and the European Union, in the framework of the RRF-2.3.1-21-2022-00004 Artificial Intelligence National Laboratory project; by the European Union's Horizon 2020 research and innovation programme under Grant Agreement no. 951832; by the V-SHIFT "Lendület" research project of the Hungarian Academy of Sciences. M. Sebők, Ł. Wordliczek and D. Stolicki acknowledge the support of the Jagiellonian University Excellence Program, DigiWorld Priority Research Area and QuantPol project. The article has relied on data collected under the Polish Ministry of Science grant no. 0395/DLG/2018/10 and the Hungarian Comparative Agendas Project (https://cap.tk.hu). We would like to thank Krisztián Boros for his help with the Transformers library. We would like to thank Zoé Baumgartner, Rafał Bieńczak, Agnieszka Karoń, Viktor Kovács, Richárd Lehoczki, Anna Sroka, Aleksandra Wójcik, and Krzysztof Ziomek for their excellent research assistance.

References

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. G. (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, *16*(1), 1–18.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473*.

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated Text Classification of News Articles: A Practical Guide. *Political Analysis*, *29*(1), 19–42.

Béchara, H., Herzog, A., Jankin, S., & John, P. (2021). Transfer learning for topic labeling: Analysis of the UK House of Commons speeches 1935–2014. *Research & Politics*, *8*(2).

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. *arXiv:2004.05150*.

Bhatia, S., Lau, J. H., & Baldwin, T. (2016). Automatic Labelling of Topics with Neural Embeddings. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, (1), 953–963.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., … Zampieri, M. (2016). Findings of the 2016 Conference on Machine Translation. *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 131–198.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258*.

Boydstun, A. E. (2013). *Making the News: Politics, the Media, and Agenda Setting*. University of Chicago Press.

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

Cardie, C., & Wilkerson, J. (2008). Text Annotation for Political Science Research. *Journal of Information Technology & Politics*, *5*(1), 1–6.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv:1406.1078*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. *arXiv:1911.02116*.

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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*.

Ding, M., Zhou, C., Yang, H., & Tang, J. (2020). CogLTX: Applying BERT to Long Texts. *Advances in Neural Information Processing Systems*, *33*, 12792–12804.

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., & Smith, N. (2020). Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. *arXiv:2002.06305*.

Dun, L., Soroka, S., & Wlezien, C. (2021). Dictionaries, Supervised Learning, and Media Coverage of Public Policy. *Political Communication*, *38*(1-2), 140–158.

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, (1), 513–520.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297.

Hillard, D., Purpura, S., & Wilkerson, J. (2008). Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. *Journal of Information Technology & Politics*, *4*(4), 31–46.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. *arXiv:1503.02531*.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *arXiv:2004.09095*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. *arXiv:2001.08361*.

Kim, Y., & Rush, A. M. (2016). Sequence-Level Knowledge Distillation. *arXiv:1606.07947*.

Kłeczek, D. (2020). Polbert: Attacking Polish NLP Tasks with Transformers. *Proceedings of the PolEval 2020 Workshop*, 79–88.

Latif, S., Rana, R., Younis, S., Qadir, J., & Epps, J. (2018). Transfer Learning for Improving Speech Emotion Classification Accuracy. *Interspeech 2018*, 257–261.

Licht, H. (2023). Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings. *Political Analysis*, *31*(3), 366–379.

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

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Loftis, M. W., & Mortensen, P. B. (2020). Collaborating with the Machines: A Hybrid Method for Classifying Policy Documents. *Policy Studies Journal*, *48*(1), 184–206.

McCann, B., Keskar, N. S., Xiong, C., & Socher, R. (2018). The Natural Language Decathlon: Multitask Learning as Question Answering. *arXiv:1806.08730*.

Monroe, B. L., & Schrodt, P. A. (2008). Introduction to the Special Issue: The Statistical Analysis of Political Text (2017/01/04). *Political Analysis*, *16*(4), 351–355.

Montgomery, J. M., Hollenbach, F. M., & Ward, M. D. (2012). Improving Predictions using Ensemble Bayesian Model Averaging (2017/01/04). *Political Analysis*, *20*(3), 271–291.

Nemeskey, D. M. (2020). *Natural Language Processing Methods for Language Modeling* [Doctoral dissertation, Eötvös Loránd University].

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.

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

Perry, C. (2013). Machine Learning and Conflict Prediction: A Use Case. *Stability: International Journal of Security and Development*, *2*(3), 56.

Peterson, A., & Spirling, A. (2018). Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems. *Political Analysis*, *26*(1), 120–128.

Pikuliak, M., Šimko, M., & Bieliková, M. (2021). Cross-lingual learning for text processing: A survey. *Expert Systems with Applications*, *165*.

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is Multilingual BERT? *arXiv:1906.01502*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(1), 5485–5551.

Ranathunga, S., Lee, E.-S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., & Kaur, R. (2023). Neural Machine Translation for Low-resource Languages: A Survey. *ACM Computing Surveys*, *55*(11), 1–37.

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

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv:1910.01108*.

Schrodt, P. A. (1990). Predicting Interstate Conflict Outcomes Using a Bootstrapped ID3 Algorithm. *Political Analysis*, *2*, 31–56.

Schrodt, P. A. (1991). Prediction of Interstate Conflict Outcomes Using a Neural Network. *Social Science Computer Review*, *9*(3), 359–380.

Sebők, M., Gajduschek, G., & Molnár, C. (2020). *A magyar jogalkotás minősége: Elmélet, mérés, eredmények (in English: The quality of Hungarian legislation: Theory, measurement, results.)* Gondolat Kiadó – Társadalomtudományi Kutatóközpont – MTA Kiválósági Kutatóhely.

Sebők, M., & Kacsuk, Z. (2021). The Multiclass Classification of Newspaper Articles with Machine Learning: The Hybrid Binary Snowball Approach. *Political Analysis*, *29*(2), 236–249.

Sebők, M., Kacsuk, Z., & Máté, Á. (2021). The (real) need for a human touch: Testing a human–machine hybrid topic classification workflow on a New York Times corpus. *Quality & Quantity*, *56*(5), 3621–3643.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? *China National Conference on Chinese Computational Linguistics*, 194–206.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems 27 (NIPS 2014)*.

Theocharis, Y., & Jungherr, A. (2020). Computational Social Science and the Study of Political Communication. *Political Communication*, *38*(1-2), 1–22.

Thrun, S. (1998). Lifelong Learning Algorithms. In S. Thrun & L. Pratt (Eds.), *Learning to Learn* (pp. 181–209). Springer New York, NY. https://doi.org/10.1007/978-1-4615-5529-2_8

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 5998–6008.

Wilkerson, J., & Casas, A. (2017). Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. *Annual Review of Political Science*, *20*, 529–544.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771*.

Wu, S., & Dredze, M. (2019). Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. *arXiv:1904.09077*.

Yang, Z., Salakhutdinov, R., & Cohen, W. W. (2017). Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks. *arXiv:1703.06345*.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 19–27.

Appendix A Transformer robustness checks

We supplemented the robustness checks regarding how various training data forms (512 tokens, extractive summary, sliding window) affect LLMs for the Polish dataset with an analysis of the Hungarian dataset. The results in Table A1 show that for the translated Hungarian data, the difference in F1 scores is relatively minor, with the sliding window approach yielding the best macro F1 scores (with the exception of RoBERTa). The 512 tokens approach comes in second for all models except for RoBERTa, which performs best with this input. The largest gap is for the original Hungarian data, where the sliding window approach yields a 0.9 performance as opposed to the 512-token input's 0.83. Interestingly, the extractive summary only outperforms the 512-token input version for the original Hungarian language data, once again pointing towards the crucial role of the front segments of bills and laws in predicting their policy topic.

Table A1: Robustness checks with Hungarian training data (macro F1 score)

Model	512 tokens	Extractive summary	Sliding window
BERT-Base (Translated Hungarian data)	0.79	0.69	0.83
BERT-Base-Hungarian (Original Hungarian data)	0.83	0.88	0.90
BERT-Large (Translated Hungarian data)	0.81	0.75	0.83
DistilBERT (Translated Hungarian data)	0.78	0.71	0.83
RoBERTa (Translated Hungarian data)	0.82	0.76	0.79

As additional robustness checks for the Transformer models, in Table A2, we focus on testing model performance with various data balancing strategies. For this test, all of the training data is translated to English from Hungarian. We tested the BERT-large model with completely balanced down-sampled training data, augmented training data where we added items to the smaller categories from the extractive summaries, and down-sampled data where the categories larger than the mean category size were truncated to the mean. The results show that BERT performance is robust to fine-tuning data composition. While the best-performing model is fine-tuned on the balanced dataset, the macro F1 score gain compared to the imbalanced model is just 0.02.

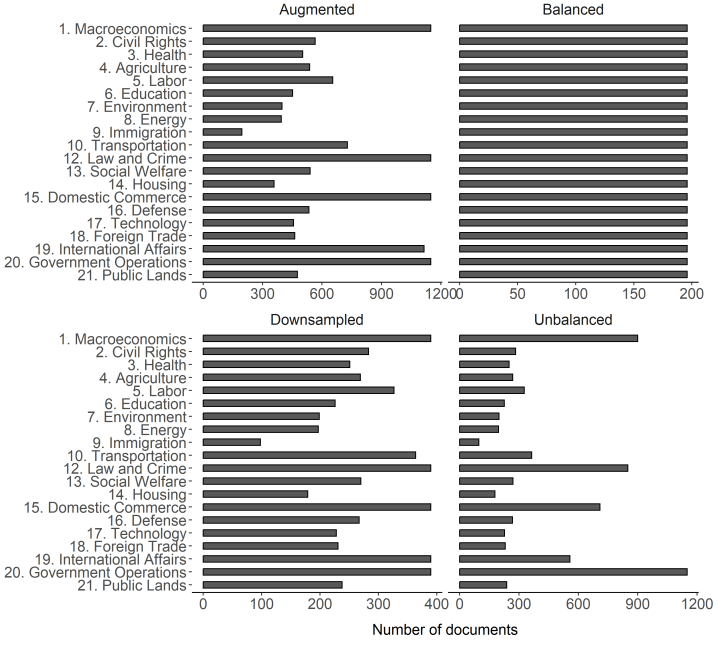
For the category distributions for each variant, see Figure A1. As the figure shows, both the augmented and down-sampled versions show improvements in terms of the more even distribution of topics over the unbalanced

Table A2: Impact of data balancing strategies on model performance

Training data	Accuracy	Macro F1 score	N
Augmented	0.83	0.83	12970
Balanced	0.85	0.84	3920
Down-sampled	0.81	0.81	5577
Imbalanced	0.83	0.82	7794

dataset. However, the gap between smaller and larger categories is still retained due to the original data's highly unbalanced nature. The trade-off made for creating the perfectly balanced sample yields only 3920 documents (as opposed to the 7794 observations in the imbalanced version).

Figure A1: Major topic distribution in different training data variants



Appendix B Translation robustness tests

For the translation-based workflows, we used the Opus-MT translation model, which itself is based on the Transformer architecture. To benchmark the performance of the model (and provide a wider context in terms of applicable machine translation resources), we translated 98 Hungarian documents to English using Opus-MT and two widely used alternatives, Google Translate and DeepL.

The average similarity values (as measured in cosine similarity) are presented in Table B1. According to the mean cosine similarity, the Opus-MT model is slightly closer to the DeepL translations but achieves comparable similarity to the Google Translate outputs as well (in both cases with a similarly minor standard deviation). Based on this evidence, the chosen open-source machine translation model's performance tracks well those of the widely used commercial translation services while offering the advantages of an open-source project which can be deployed at no cost.[17] These results further buttress the viability of the proposed workflow for low-resource languages and for comparative projects with limited funding.

Table B1: Comparing translation models

Translation engine pair	Mean similarity	Std.dev	Minimum	Maximum	N
Opus-MT x DeepL	0.93	0.05	0.52	0.99	98
Opus-MT x Google	0.92	0.04	0.77	0.99	98

[17] For more information, the performance of the Opus-MT model on various benchmark datasets can be checked here: https://opus.nlpl.eu/leaderboard/.