Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 5.2 (2023) 1–20
HTTPS://DOI.ORG/10.5117/CCR2023.2.3.HO

Evaluating Transferability in Multilingual Text Analyses

Justin Chun-ting Ho
Communication Data and Network Analytics Lab, Academia Sinica, Taiwan

Chung-hong Chan
GESIS - Leibniz-Institut für Sozialwissenschaften, Germany

Abstract
Multilingual text analysis is increasingly important to address the current narrow focus of English and other Indo-European languages in comparative studies. However, there has been a lack of a comprehensive approach to evaluate the validity of multilingual text analytic methods across different language contexts. To address this issue, we propose that the validity of multilingual text analysis should be studied through the lens of transferability, which assesses the extent to which the performance of a multilingual text analytic method can be maintained when switching from one language context to another. We first formally conceptualize transferability in multilingual text analysis as a measure of whether the method is equivalent across language groups (linguistic transferability) and societal contexts (contextual transferability). We propose a model-agnostic approach to evaluate transferability using (1) natural and synthetic data pairs, (2) manual annotation of errors, and (3) the Local Interpretable Model-Agnostic Explanations (LIME) technique. As an application of our approach, we analyze the transferability of a multilingual BERT (mBERT) model fine-tuned with annotated manifestos and media texts from five Indo-European language-speaking countries of the Comparative Agendas Project. The transferability is then evaluated using natural and synthetic parliamentary data from the UK, Basque, Hong Kong, and Taiwan. Through the evaluation of transferability, this study sheds light on the common causes that lead to prediction errors in multilingual text classification using mBERT.

Keywords: Multilingual Text Analysis, Topic Classification, Transfer Learning, Error Analysis, Large Language Model

Baden et al. (2022) observe an "English before everything" tendency in computational text analysis where most of the current methods and applications focus only on English. There is an urgent need to address this myopic vision. Unfortunately, current methodological research on multilingual text analysis still predominantly centers around various Indo-European languages (e.g., Dobbrick et al., 2021; Glavaš et al., 2017; Reber, 2019), with limited inclusion of non-Indo-European languages, except for instances such as Maier et al. (2022). This narrow focus has an important implication for analyzing text for political communication research: only four out of twenty full democracies do not use Indo-European languages (Finland, Taiwan, South Korea, and Japan) while only three authoritarian countries use Indo-European languages. The heavy focus on Indo-European languages precludes what texts from which political systems one can analyze.

Recent advancement in large language models (LLMs) offers a promising solution to multilingual text analysis. BERT (Bidirectional Encoder Representations from Transformers), a language model developed by Google (Devlin et al., 2018), is an exemplar model. Despite the high performance, these models often require large amounts of data and computing power to train, which might not always available for many researchers. To deal with this problem, researchers turn to transfer learning, a technique to apply knowledge learned in one domain to complete tasks of another domain (Azunre, 2021). One popular approach is to leverage annotated text compiled by international data curation and annotation teams, such as the Manifesto Project and Comparative Agendas Project (CAP) (John et al., 2013; Merz et al., 2016), to fine-tune pre-trained language models. Previous work demonstrates the potential of this approach for classifying political texts in policy domain (Koh et al., 2021; Terechshenko et al., 2020). In addition to monolingual work, recent work also makes use of mBERT, a multilingual version of BERT, to classify multilingual texts. While preliminary evidence demonstrates mBERT's performance for non-Indo-European languages (Chan et al., 2020; Wu & Dredze, 2019), its applicability in downstream tasks needs to be validated. In particular, we need to evaluate in what extend and, more importantly, **why** does transfer learning work. However, this task is notoriously difficult, not least because of the "black box" nature of these language models.

The current paper attempts to shed light into these black boxes by first conceptualize **transferability**. Second, we propose an approach to evaluate the transferability of multilingual language models despite their inherent uninterpretability. Finally, we demonstrate how to conduct the evaluation of transferability and present a suggested workflow.

Transferability

We conceptualize transferability as the extend to which the performance of a (multilingual) text analytic method can be maintained when switching from one language to another. For example, if a text model trained on English parliamentary text data can achieve same performance in Chinese parliamentary text, this method is said to have high transferability from English to Chinese for this particular task. From the perspective of the model, we refer to English and Chinese as seen and unseen languages respectively. Model performance can be measured both quantitatively, using traditional machine learning metrics such as precision and recall, or qualitatively, through the analysis of misclassified cases as in communication science (Van Atteveldt et al., 2021).

However, this conceptualization can be easily confused with a mere language transfer problem, specifically *What is the performance of the model trained on text data in some seen languages for text data in an unseen language?*. This question primarily addresses the notion of *linguistic transferability*. Although this question is significant, it has been observed that the mismatch of context or genre can also impact model performance even in monolingual analyses (Koh et al., 2021; Osnabrügge et al., 2021; Terechshenko et al., 2020). This problem becomes more pronounced in multilingual analysis in social sciences, where the text data used for fine-tuning a pretrained model encompasses both linguistic and contextual information (as demonstrated by the correlation between democracy and Indo-European languages mentioned earlier), with the latter being tacit and often overlooked. As a simple example, an English speaker at the UK Parliament might refer to "London" as the "capital", but a Chinese speaker at the Taiwan Legislative Yuan would probably not. An additional question to ask when evaluating transferability is: *What is the performance of the model trained on text data from the contexts of seen languages when applied to text data from the context of an unseen language?*. This additional question address the *contextual transferability*.

With this two-dimensional conceptualization of transferability, we propose a mixed-method approach to operationalize transferability. In this paper, we use a fine-tuned mBERT model as a case study to demonstrate how to evaluate transferability. However, it is important to note that our operationalization is model-agnostic (See Online Appendix).[1]

[1]Accessible at https://osf.io/fdcea/?view_only=6f683097162d4ef582ba0be63d026de5; while this article focuses on mBERT, we have included an additional example to demonstrate how to apply the workflow for Support Vector Machine using *scikit-learn*.

Case study: Fine-tuned mBERT for topical classification

The task of classifying topics in text data is a heavily-studied task in the methodological literature of communication research. For example, methods such as latent Dirichlet allocation topic modeling and supervised machine learning models have achieved various success (Maier et al., 2018). Previous attempts to analyse topics in multilingual text data include machine translation (de Vries et al., 2018; Lind et al., 2021; Lucas et al., 2015; Maier et al., 2022; Reber, 2019), multilingual dictionary construction (Maier et al., 2022), and aligned word embeddings (Chan et al., 2020).

For this task, we used the annotated texts from five countries of CAP (UK, Germany, France, Italy, and Spain).[2] The following topics were annotated: 1) Macroeconomics, 2) Civil Rights, Minority Issues, and Civil Liberties, 3) Health, 4) Agriculture, 5) Labour and Employment, 6) Education, 7) Environment, 8) Energy, 9) Immigration, 10) Transportation, 12) Law, Crime, and Family Issues, 13) Social Welfare, 14) Community Development, Planning and Housing Issues, 15) Banking, Finance, and Domestic Commerce, 16) Defence, 17) Space, Science, Technology and Communications, 18) Foreign Trade, 19) International Affairs and Foreign Aid, and 21) Public Lands, Water Management, Colonial and Territorial Issues.[3]

mBERT

mBERT is an LLM trained on a massive amount of multilingual Wikipedia data in 104 languages without much human annotation and advertised to provide (linguistic) transferability of performance across languages (Pires

[2]For UK, we used party manifestos and media texts (collected by John et al. (2013)); For Germany, we used party manifestos (collected by Breunig et al. (2021)); For France, we used party manifestos and government communications (collected by Emiliano Grossman, Sylvain Brouard, Isabelle Guinaudeau, Caterina Froio, Tinette Schnatterer, and Simon Persico); For Italy, we used party manifestos (collected by Borghetto et al. (2019)); For Spain, we used manifestos and media texts (collected by Laura Chaqués-Bonafont, Anna M. Palau and Luz M. Muñoz, with the collaboration of graduate students and the financial support of the Spanish Ministry of Innovation and Science and the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR). Neither these public institutions nor the original collectors of the data bear any responsibility for the analysis reported here.)

[3]Due to the nature of our prediction data, texts from parliamentary material, the CAP topic category of "Government Operations" was excluded as preliminary analysis shows that the linguistic features of parliamentary material tend to drive all predictions towards the said category in which "Parliamentary Operations" is included in one of the subcategory. Second, a category of "No Topic" was introduce to capture boilerplates and texts that carry no substantive meaning.

et al., 2019). It has been demonstrated to have zero-to-few-shot cross-lingual transfer potential: an approach to conduct multilingual analysis by fine-tuning a model in a source language, often a high-resource language, and making predictions on other languages (Pires et al., 2019; Wu & Dredze, 2019). Cross-lingual transfer is especially important for non-Indo-European languages, where annotated training data are often hard to obtain. Researchers have also attempted to extend mBERT to languages beyond the original 104 languages and shown promising capacities (Wang et al., 2020). Real-life evaluations, however, indicate that the downstream performance is significantly better for Indo-European languages comparared to other languages. Wu and Dredze (2020), for instance, show that while mBERT performs well on European languages such as English and German, its performance is notably worse for Asian languages such as Chinese and Japanese.

Transfer leaning has been applied for topic classification in monolingual settings such as English (Terechshenko et al., 2020) and German (Widmann & Wich, 2022) by fine-tuning. Fine-tuning refers to the procedure of providing a small amount of training data to an LLM to enable it to perform a specific task, for example topic classification. The approach can also be extended to multilingual text corpora. Guo et al. (2022), for instance, propose a mBERT-based tool for multilingual content analysis, albeit the tool has only been tested with English and German data.

Fine-tuning

We began by fine-tuning the off-the-shelf mBERT model, a neural network with 12 layers. We initialized the model with pre-trained parameters, froze the first 8 layers of the model, and trained the last 4 layers with the five-country annotated texts from CAP (thereafter, fine-tuning data). Freezing initial layers is a common technique used in transfer learning (Azunre, 2021). The intuition behind is to preserve the knowledge the model has previously learned while allowing room for calibration based on new input.

Quantitative Evaluation of performance in the traditional sense

The first step after the fine-tuning is to evaluate. Evaluation usually refers to the assessment of predictive performance using metrics such as correct classification rate (CCR), sensitivity, specificity, and F-score. While the predictive performance of a language model's downstream prediction tasks is often the focus of computer science research, it serves as a a means rather

than an end for social science applications as long as they pass validity tests (Grimmer et al., 2021; Waldherr et al., 2021). Dobbrick et al. (2021) refer to these applications as "shotgun approach", where users lack insight into why these models work.. Nevertheless, the first step of our transferability evaluation still relies on the evaluation of predictive performance.

We used CCR (total number of correctly classified cases divided by total of cases) in this analysis (Cross tabulations are provided in the Online Appendix).

Qualitative Evaluation of performance: Error analysis

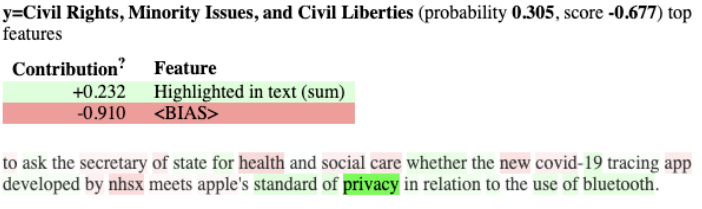
Apart from predictive performance, it is also important to explain why a language model is (not) accurate. While interpretability is currently a new requirement of machine learning systems (Lipton, 2018), LLMs are inherently uninterpretable due to the complexity of the model architecture. For instance, it is not possible to examine the regression coefficient of a specific feature (word) as in a logistic regression model. To gain insights into the black box nature of LLMs, we employ (qualitative) error analysis.

Traditionally, this process is done by qualitative reading of misclassified cases (Van Atteveldt et al., 2021). We propose an improvement that incorporates the Local Interpretable Model-Agnostic Explanations (LIME) technique (Ribeiro et al., 2016) to identify the words that contributed to the prediction error. LIME involves randomly perturbing each input text to examine how the prediction changes. By comparing the predicted probabilities across the changes, we can infer the relative importance of each word to the final prediction, thus making it interpretable.

Figure 1 is a visualization of LIME. The model predicts that the input text (the lowest part) is in the topic *Civil Rights, Minority Issues, and Civil Liberties*. The words in the text are highlighted in different colors and in different shades to represent a word's level of influence and its directions (green: for, red: against) to the model prediction. For example, the word "privacy" is highlighted in bright green, suggesting it leads the model heavily to the decision of *Civil Rights, Minority Issues, and Civil Liberties*. Red words like "health" and "app" on the other hand lead the model to predict other topics.

By looking at the words that contribute to an incorrect prediction, a hypothesis for the classification error can be formed. The result of the error analysis can then be used to understand the bias and shortcomings of the model for the classification task.

Figure 1: LIME Representation (Attention Error)



Evaluation of transferability with natural and synthetic data pairs

Due to the aforementioned blackbox nature of LLMs, we can only evaluate transferability using the **model-based testing** technique (Bringmann & Kr, 2008). We gain knowledge about the model by supplying carefully crafted input data to the multilingual model and assess its transferability by by comparing the model's result to the expected outputs.

Base on our conceptualization of transferability, the seen languages of our fine-tuned mBERT model are English, German, French, Italian, and Spanish. The unseen languages are Chinese and Basque. The seen societal contexts are UK, Germany, France, Italy, and Spain and the unseen societal contexts are Taiwan and Hong Kong. We derived eight cases chosen along two dimensions (see Table 1). The first dimension is linguistic similarity (LS) for the evaluation of linguistic transferability: a language is said to have high LS if the *language* is seen by the model during the fine-tuning (English, French, German, Italian, and Spanish) and vice versa. The second dimension is contextual similarity (CS) for the evaluation of contextual transferability: a societal context is said to have high CS if the *country* of the texts is seen by the model during the fine-tuning (UK, France, Germany, Italy, and Spain) and vice versa.

Natural cases and synthetic cases

For each case, parliamentary material such as meeting titles, summaries, and written questions were collected with its official Application Programming Interface or data portal.[4]

There are five natural and three synthetic cases (see Table 1). UK is selected as a case for "high LS, high CS" since both the language (English)

[4]For UK, we use the Member of Parliament's written questions; for Taiwan, we use the Executive Yuan's reply; for Hong Kong, we use the meeting summary of Legislative Council Panels; for Basque, we use the title of the initiatives of Basque Parliament.

and the societal context (UK) are seen by the model. Taiwan and its official language, Traditional Chinese, are both unseen language and context, so it is selected as a case for "low LS, low CS". Since Hong Kong offers both data in English and Chinese (both of them are official languages of Hong Kong), they are selected as natural cases of "high LS, low CS" and "low LS, low CS" respectively. The Basque Autonomous Community and its parliament are located in the seen context of Spain but the Basque language is an unseen language. Thus, it is selected as a natural case of "low LS, high CS". We decided to use Basque to ensure all of the unseen languages (Basque and Chinese) are non-Indo-European languages, while all of the of seen languages (English, German, French, Italian, and Spanish) are Indo-European languages.[5]

We also created three synthetic cases by translating (1) the Taiwan and Basque material to English to artificially modify them from low LS to high LS and (2) the UK material to Chinese to artificially modify it from high LS to low LS. The Basque material was translated by a professional translator hired on Upwork, a freelancing platform. The UK and Taiwan material was translated by one of the authors, who is a fluent speaker of both languages. These synthetic cases allow direct attribution of the performance impact based on the two dimensions of transferability. For example, the variation in predictive performance observed when transitioning from Taiwan's original content in Traditional Chinese to translated content in English can be directly attributed to linguistic differences, as the context remains constant; while the difference in predictive performance between the UK English content and the Taiwan content in English can be attributed to contextual differences.

Table 1: Case Selection

	High LS (Seen)	Low LS (Unseen)
High CS (Seen)	UK (English), Basque (English)	Basque (Basque), UK (Chinese)
Low CS (Unseen)	Hong Kong (English), Taiwan (English)	Hong Kong (Chinese), Taiwan (Chinese)

CS: Contextual Similarity; LS: Linguistic Similarity

[5]We decided not to use some other possible "low LS, high CS" cases such as Catalonia (Catalan) and Scotland (Scottish Gaelic) because the languages are in the Indo-European language family.

Evaluate transferability quantitatively

The first way to evaluate transferability is to compare the predictive performance metrics by cases. Figure 2 shows the CCR for each case.

To construct the validation set for evaluating the performance of the our language model, we employ stratified sampling to ensure full coverage of all topics. For each case, a sample from each predicted topic was drawn and coded according to the official CAP coding scheme by a co-author not involved in the above fine-tuning. This serves as the "gold standard" for our evaluation. The final validation set consists of 421 texts, around 50 texts per case. To test inter-coder reliability, a sub-sample of 40 texts was coded by both authors and the result shown substantial inter-coder agreement (80% agreement, Krippendorff's $\alpha = 0.772$, Cohen's $\kappa = 0.769$).

An important finding from Figure 2 (Left) is that the content from the Basque region in the original Basque language has a significantly lower predictive performance than the same content translated to English. Since the content is identical between the two Basque cases, the change in predictive performance can be directly attributed to the linguistic difference. The significant performance gap observed indicates the limited linguistic transferability of our model from Basque to English. A similar pattern emerges for Hong Kong and the UK, where the model performs better on English text compared to Chinese text. A peculiar case is observed with translated material from Taiwan, which showed an opposite trend. The implication of this will be discussed in the conclusion.

Figure 2: Correct Classification Rate by Case, Language, and Case Type

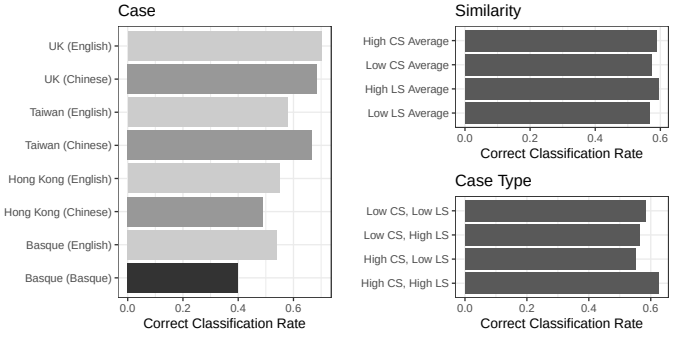


Figure 2 (Bottom Right) shows the average CCR along the dimensions of linguistic and contextual similarity. Once again, we observe that the "high LS, high CS" cases exhibit the highest performance. However, it is important

to interpret the remaining results cautiously since Basque and Chinese are grouped together. Despite both languages being non-Indo-European, their performance in this task differs significantly. While Basque material (0.40) has the lowest CCR, there are negligible difference between English (0.60) and Chinese (0.62).

Evaluate transferability qualitatively

Another approach to evaluate transferability is through qualitative error analysis (Van Atteveldt et al., 2021). In our study, one of the co-authors conducted error analysis by examining and annotating all misclassified texts. Our approach introduces two innovations beyond the traditional qualitative reading method.

First, a LIME representation was generated for each misclassified text using the *eli5* python package (eli5 developers, 2022). Second, our synthetic approach allows us to directly compare the LIME representations of a misclassified text in its original language and translated language. This comparison allows us to form better hypotheses about the possible reasons for the prediction error of each misclassified text. We refer to this process as **error annotation**.

Table 2 shows the result of the error annotation. Our analysis suggests that the possible reasons for prediction errors in this study can be categorized into eight categories. These fine-grained hypotheses for prediction errors were only made possible by incorporating the two innovations in error annotation.

Attention Error

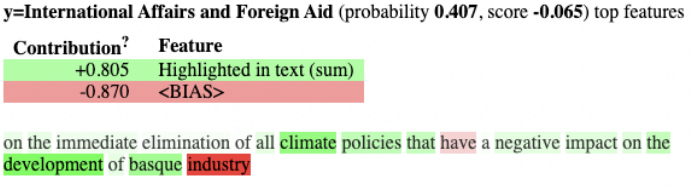
Attention Error refers to cases where the prediction is influenced by a word that is not the main focus of the text, but the influence on the prediction by the influential word is correct. An example illustrating this is depicted in Figure 1, where the text primarily focuses on the development of a new mobile phone application, which should be categorized as *Space, Science, Technology, and Communications*. However, the incorrect prediction is influenced by the word "privacy", which correctly influences the prediction towards *Civil Rights, Minority Issues, and Civil Liberties*.

Model Deficiency

Model Deficiency refers to the cases when the prediction is influenced by a word that is the main focus of the text, but the influence on the prediction

by the influential word is wrong. For example, in Figure 3 the word "climate" contributed to the prediction of *International Affairs and Foreign Aid* word, while the word should be categorized as *Environment*.

Figure 3: LIME Representation (Model Deficiency)



Difficult Topic

Difficult Topic refers to the texts that belong to topics that cannot be easily distinguished from one and other. Example pairs are (1) *Macroeconomics* versus *Banking, Finance and Domestic Commerce* and (2) *Foreign Trade* versus *International Affairs and Foreign Aid*.

Ngram Error

Ngram Error refers to the cases where the prediction is influenced by a single word, but the word would fall into a different topic when combined with another neighboring word. For example, "food" should usually be categorized as *Agriculture* but "food business" should be categorized as *Banking, Finance and Domestic Commerce*.

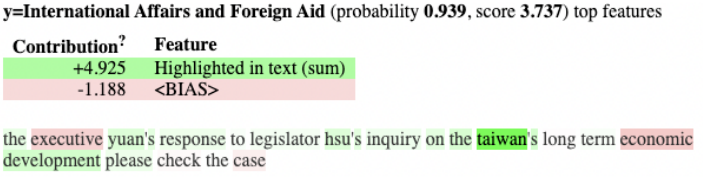
Context Error

Context Error refers to the cases where the prediction would be correct if the text is understood within another societal context. For example, in Figure 4 the word "Taiwan" was categorized as *International Affairs and Foreign Aid* in the original training data, which would only be correct if the text is not from Taiwan.

Language Misinterpretation

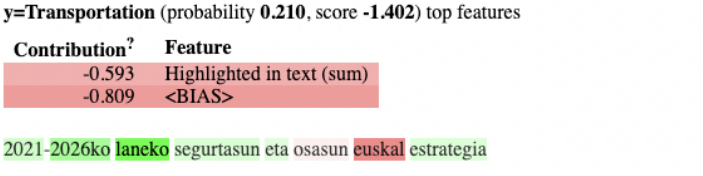
Language Misinterpretation refers to the cases where the model misunderstood the text. For instance, "laneko" (means "at work" in Basque) is

Figure 4: LIME Representation (Context Error)



predicted as *Transportation* in one case, potentially because of its similarity with "lane" in English (Figure 5). This issue is specific to transformer models such as mBERT because they use a common feature space across all languages.

Figure 5: LIME Representation (Language Misinterpretation)



Topic Unknown and Missing Topic

Topic Unknown refers to the cases where the topic of the text cannot be determined in accordance with the CAP codebook. For example, one text reads "Present the assessment of the last two years", which does not provide any substantial information about the specific policy area. Missing Topic refers to the cases when the topic can be determined, but the model does not have a specific category for that particular topic.. For instance, since the topic of "Government Operations" was removed during model training, all texts belong to this topic will fall into this category.

Cross-language errors

In additional to examining the errors in isolation, our research design allows us to to analyze errors based on language pairs. Three types of errors can be identified: errors occurring in both languages, errors only in English, and errors only in non-English languages. The percentages for each error type are presented in Table 3, while Figure 6 provides a closer examination of the potential causes of these errors. It is observed that Difficult Topic affects

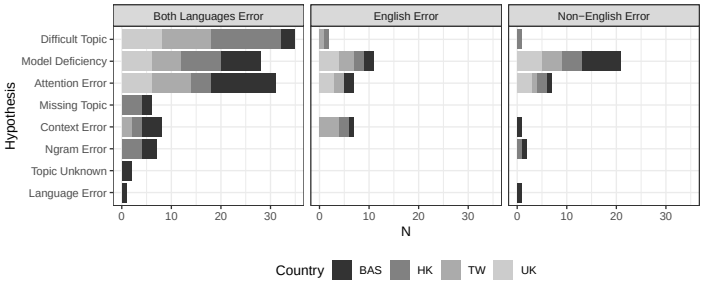
Table 2: Error Annotation

Hypothesis	High CS High LS	High CS Low LS	Low CS High LS	Low CS Low LS	Total
Attention Error	31.67%	30.00%	18.64%	18.64%	24.79%
Context Error	6.67%	6.67%	13.56%	13.56%	10.08%
Difficult Topic	8.33%	10.00%	25.42%	25.42%	17.23%
Language Error	3.33%	1.67%	0.00%	0.00%	1.26%
Missing Topic	1.67%	1.67%	3.39%	3.39%	2.52%
Model Deficiency	41.67%	45.00%	33.90%	33.90%	38.66%
Ngram Error	5.00%	3.33%	5.08%	5.08%	4.62%
Topic Unknown	1.67%	1.67%	0.00%	0.00%	0.84%

Percentages are computed by dividing the counts for an individual cell by the total count for the column; CS: Contextual Similarity; LS: Linguistic Similarity.

predictions in both languages, followed by Attention Error, albeit to a lesser extent. Model Deficiency appears to have a greater impact on non-English predictions, while Context Error is more prominent in English predictions.

Figure 6: Cross-language errors



Conclusion

Figure 7 outlines our proposed workflow for evaluating transferability. The key innovations of our approach are the emphases on qualitative evaluation and synthetic cases. A practical guide along with codes to replicate the workflow are included in the Online Appendix.

Using our two-dimensional conceptualization of transferability, our evaluation provides researchers with a deeper understanding of the performance of multilingual classification models compared to traditional approaches

Table 3: Cross-Language Error

Error Type	BAS	HK	TW	UK	Total
Both Languages Error	42.00%	37.25%	31.58%	19.30%	32.09%
English Error	8.00%	9.80%	17.54%	12.28%	12.09%
Non-English Error	20.00%	15.69%	8.77%	14.04%	14.42%
No Error	30.00%	37.25%	42.11%	54.39%	41.40%

Percentages are computed by dividing the counts for an individual cell by the total count for the column.

that solely focus on predictive performance. This is clearly demonstrated in our case study, where we evaluated a fine-tuned mBERT model using the framework of transferability. The evaluation was made possible by the two unique features of our approach.

Synthetic cases

The idea of translating content is not new and has been used several times in the methodological literature, mostly as an enabler of other classification tasks (de Vries et al., 2018; Reber, 2019). Our proposal is to ultise translation to modify LS while keeping CS constant. We show that the model on average has a 3-percent (absolute, same below) penalty in CCR when moving from high LS to low LS, while there was only 2-percent penalty when moving from high CS to low CS (Figure 2).

We observe a 5.9-percent difference in CCR between English and Chinese in the Hong Kong case. The translation of UK English material to Chinese decreased the CCR by 1.8 percent while the translation from Basque to English increased the CCR by 14 percent. The results shows the limited linguistic transferability of mBERT *ceteris paribus*.

Future applications of multilingual text analysis are recommended to include a routine check that involves performance comparison on both the original content and the synthetic translated content. However, it is important to note that the translations were conducted by either professional translators or fluent native speakers. The current study provides no evidence on whether or not this approach can be done with machine translation, the default approach used in many previous studies (de Vries et al., 2018; Reber, 2019). Future studies could look into this direction for generating synthetic cases.[6]

[6]We decided not to analyse machine translation in the current study since modern ma-

The importance of qualitative evaluation of transferability

Our approach places a strong emphasis on qualitative evaluation. The error annotation procedure allows us to **explain** how a multilingual model could go wrong (Table 2). While most of the hypotheses from our error annotation are not directly related to multilinguality, "Context Error" and "Language Misinterpretation" can only arise with multilingual transfer learning. These insights are only attainable through qualitative evaluation.

Our qualitative evaluation highlights the need for caution when conducting multilingual analyses, particularly in relation to these sources of errors. Additionally, our findings indicate that there is significant room for research into context-aware machine learning, for example reduce language misinterpretaion errors with a better representation of the feature space for LLMs.

Possible extensions

Resource transferability

The Taiwan case is an exception as the CCR was decreased by 8.8 percent from the original to the translated. While the mBERT model appears to perform better on English material in most cases, the average accuracy of prediction for Chinese content is comparable to that of English content with only a 2-percent difference.

The high performance of mBERT for Chinese material could be explained by its original training data (not the fine-tuning data). Research on how the performance of the downstream tasks of an LLM by the training material is the central question of algorithmic bias and researchers have begun to investigate this issue (Yang & Roberts, 2021). mBERT was trained on Wikipedia data and Chinese is one of top languages of Wikipedia in terms of the amount of articles created: There are 6 million articles in English, 1.3 million in Chinese, and only 400 thousand in Basque. The relative abundance of Chinese articles to Basque articles means that mBERT probably had enough training data to learn about the Chinese language, but not the Basque language. We encourage future work to further investigate the influence of "resourcefulness" of a language on the text analytic tasks. This aspect, admittedly, has not been conceptualized in our original conceptual-

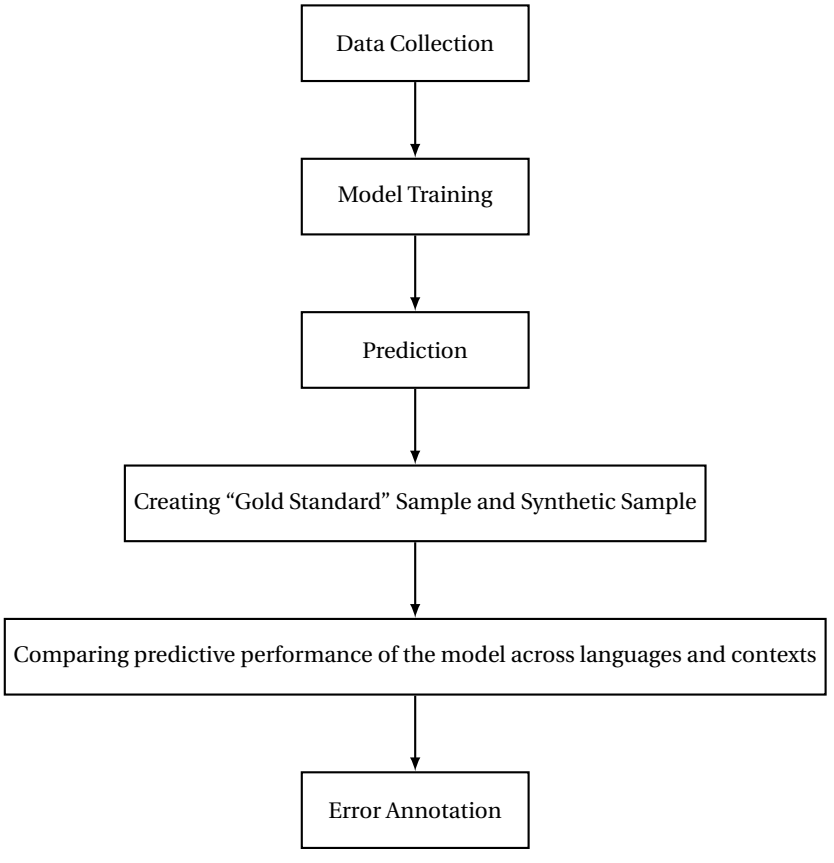
chine translation techniques are also based on LLMs. To prevent the tautological situation of evaluating a LLM by the synthetic cases generated by another LLM, we decided to use human translation. It is also in the communication science traditional to use human intervention as the gold standard (Van Atteveldt et al., 2021).

ization of transferability. We therefore propose a possible third dimension of transferability: resource transferability. For example, Chinese and English have high resource transferability, because they have the similar level of resourcefulness despite being in different language families. The same cannot be said between Chinese and Basque. Joshi et al.'s (2020) quantification of language resourcefulness is useful for determining resourcefulness. And the disparity in resourcefulness has more to do with researchers' collective attention than language family or number of speakers (Baden et al., 2022).

Similarity as a continuum

Our approach utilizes a binary measurement for linguistic and contextual similarities, categorizing a societal context as either present or not present in the fine-tuning data. However, it is important to acknowledge that this binary approach may overlook the nuanced differences between societal contexts, such as the distinction between Taiwan and Hong Kong in terms of their political systems. Lumping them together based solely on their presence in the fine-tuning data may not fully capture the contextual dissimilarities between these two places. Contextual similarities can be operationalized as a continuous measurement, for example by using survey results from the World Value Survey, and thus enabling a more granular assessment of the transferability of multilingual models.

Figure 7: Suggested Workflow



References

Azunre, P. (2021). *Transfer Learning for Natural Language Processing*. Manning.

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, *16*(1), 1–18. https://doi.org/10.1080/19312458.2021.2015574

Borghetto, E., Carammia, M., & Russo, F. (2019, March). The Italian Agendas Project. In F. R. Baumgartner, C. Breunig, & E. Grossman (Eds.), *Comparative Policy Agendas: Theory, Tools, Data* (pp. 120–128). Oxford University Press. https://doi.org/10.1093/oso/9780198835332.003.0013

Breunig, C., Guinaudeau, B., & Schnatterer, T. (2021). Policy agendas in Germany – database and descriptive insights. *The Journal of Legislative Studies*, 1–13. https://doi.org/10.1080/13572334.2021.2010395

Bringmann, E., & Kr, A. (2008). Model-Based Testing of Automotive Systems. *2008 International Conference on Software Testing, Verification, and Validation*. https://doi.org/10.1109/icst.2008.45

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

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

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. https://doi.org/10.1017/pan.2018.26

Dobbrick, T., Jakob, J., Chan, C.-H., & Wessler, H. (2021). Enhancing Theory-Informed Dictionary Approaches with "Glass-box" Machine Learning: The Case of Integrative Complexity in Social Media Comments. *Communication Methods and Measures*, 1–18. https://doi.org/10.1080/19312458.2021.1999913

eli5 developers. (2022). *Eli5*. https://eli5.readthedocs.io/en/latest/overview.html

Glavaš, G., Nanni, F., & Ponzetto, S. P. (2017). Cross-Lingual Classification of Topics in Political Texts. *Proceedings of the Second Workshop on NLP and Computational Social Science*, 42–46. https://doi.org/10.18653/v1/W17-2906

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, *24*(1), 395–419. https://doi.org/10.1146/annurev-polisci-053119-015921

Guo, L., Su, C., Paik, S., Bhatia, V., Akavoor, V. P., Gao, G., Betke, M., & Wijaya, D. (2022). Proposing an Open-Sourced Tool for Computational Framing Analysis of Multilingual Data. *Digital Journalism*, 1–22. https://doi.org/10.1080/21670811.2022.2031241

John, P., Bertelli, A. M., Jennings, W. J., & Bevan, S. (2013). *Policy agendas in British politics*. Palgrave Macmillan.

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 preprint arXiv:2004.09095*.

Koh, A., Boey, D. K. S., & Béchara, H. (2021, April). *Predicting Policy Domains from Party Manifestos with BERT and Convolutional Neural Networks* (preprint). SocArXiv. https://doi.org/10.31235/osf.io/fjh4q

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). https://doi.org/10.5117/CCR2021.3.001.LIND

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

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

Maier, D., Baden, C., Stoltenberg, D., De Vries-Kedem, M., & Waldherr, A. (2022). Machine Translation Vs. Multilingual Dictionaries Assessing Two Strategies for the Topic Modeling of Multilingual Text Collections. *Communication Methods and Measures*, *16*(1), 19–38. https://doi.org/10.1080/19312458.2021.1955845

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures*, *12*(2-3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Merz, N., Regel, S., & Lewandowski, J. (2016). The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*, *3*(2), 205316801664334. https://doi.org/10.1177/2053168016643346

Osnabrügge, M., Ash, E., & Morelli, M. (2021). Cross-domain topic classification for political texts. *Political Analysis*, 1–22. https://doi.org/10.1017/pan.2021.37

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is Multilingual BERT? [arXiv: 1906.01502]. *arXiv:1906.01502 [cs]*. Retrieved March 3, 2022, from http://arxiv.org/abs/1906.01502

Reber, U. (2019). Overcoming Language Barriers: Assessing the Potential of Machine Translation and Topic Modeling for the Comparative Analysis of Multilingual Text Corpora. *Communication Methods and Measures*, *13*(2), 102–125. https://doi.org/10.1080/19312458.2018.1555798

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Terechshenko, Z., Linder, F., Padmakumar, V., Liu, F., Nagler, J., Tucker, J. A., & Bonneau, R. (2020). A Comparison of Methods in Political Science Text Classification: Transfer Learning Language Models for Politics. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3724644

Van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis:comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 1–20. https://doi.org/10.1080/19312458.2020.1869198

Waldherr, A., Geise, S., Mahrt, M., Katzenbach, C., & Nuernbergk, C. (2021). Toward a Stronger Theoretical Grounding of Computational Communication Science: How Macro Frameworks Shape Our Research Agendas. *Computational Communication Research*, *3*(2), 1–28. https://doi.org/10.5117/CCR2021.02.002.WALD

Wang, Z., K, K., Mayhew, S., & Roth, D. (2020). Extending Multilingual BERT to Low-Resource Languages. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2649–2656. https://doi.org/10.18653/v1/2020.findings-emnlp.240

Widmann, T., & Wich, M. (2022). Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in german political text. *Political Analysis*, 1–16. https://doi.org/10.1017/pan.2022.15

Wu, S., & Dredze, M. (2019). Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 833–844. https://doi.org/10.18653/v1/D19-1077

Wu, S., & Dredze, M. (2020). Are All Languages Created Equal in Multilingual BERT? [arXiv: 2005.09093]. *arXiv:2005.09093 [cs]*. Retrieved March 3, 2022, from http://arxiv.org/abs/2005.09093
Comment: RepL4NLP Workshop 2020 (Best Long Paper).

Yang, E., & Roberts, M. E. (2021). Censorship of online encyclopedias. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3442188.3445916