Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI

Moritz Laurer, Wouter van Atteveldt, Andreu Casas and Kasper Welbers

Department of Communication Science, Vrije Universiteit Amsterdam, Amsterdam, Netherlands. Email: m.laurer@vu.nl, wouter.van.atteveldt@vu.nl, a.casassalleras@vu.nl, k.welbers@vu.nl

Abstract

Supervised machine learning is an increasingly popular tool for analyzing large political text corpora. The main disadvantage of supervised machine learning is the need for thousands of manually annotated training data points. This issue is particularly important in the social sciences where most new research questions require new training data for a new task tailored to the specific research question. This paper analyses how deep transfer learning can help address this challenge by accumulating "prior knowledge" in language models. Models like BERT can learn statistical language patterns through pre-training ("language knowledge"), and reliance on task-specific data can be reduced by training on universal tasks like natural language inference (NLI; "task knowledge"). We demonstrate the benefits of transfer learning on a wide range of eight tasks. Across these eight tasks, our BERT-NLI model fine-tuned on 100 to 2,500 texts performs on average 10.7 to 18.3 percentage points better than classical models without transfer learning. Our study indicates that BERT-NLI fine-tuned on 500 texts achieves similar performance as classical models trained on around 5,000 texts. Moreover, we show that transfer learning works particularly well on imbalanced data. We conclude by discussing limitations of transfer learning and by outlining new opportunities for political science research.

Keywords: machine learning, computational methods, text as data, transfer learning

1. Introduction

From decades of political speeches to millions of social media posts – more and more politically relevant information is hidden in digital text corpora too large for manual analyses. The key promise of computational text analysis methods is to enable the analysis of these corpora by reducing the need for expensive manual labor. These methods help researchers extract meaningful information from texts through algorithmic support tools and have become increasingly popular in political science over the past decade (Benoit 2020; Grimmer and Stewart 2013; Lucas *et al*. 2015; Van Atteveldt, Trilling, and Calderon 2022; Wilkerson and Casas 2017).

Supervised machine learning is one such algorithmic support tool (Osnabrügge, Ash, and Morelli 2021). Researchers manually create a set of examples for a specific task (training data) and then train a model to reproduce the task on unseen text. The main challenge of this approach is the creation of training data. Supervised models require relatively large amounts of training data to obtain good performance, making them a "nonstarter for many researchers and projects" (Wilkerson and Casas 2017). Lack of data is particularly problematic in the social sciences where most new research questions entail a new task (task diversity) and some concepts of interest are only present in a small fraction of a corpus (data imbalance). Compared to the natural language processing (NLP) literature, for example, political scientists are less interested in recurring benchmark tasks with rich and artificially balanced data. The ensuing data scarcity problem is probably an important reason for the greater popularity of unsupervised approaches in the social sciences.

***Political Analysis* (2024)**
vol. 32: 84–100
DOI: 10.1017/pan.2023.20

Published
9 June 2023

Corresponding author
Moritz Laurer

Edited by
Jeff Gill

Unsupervised approaches are difficult to tailor to specific tasks and are harder to validate, but they do not require training data (Denny and Spirling 2018; Miller, Linder, and Mebane 2020, 4).

This paper argues that this data scarcity problem of supervised machine learning can be mitigated through deep transfer learning. The main assumption of transfer learning is that machine-learning models can learn "language knowledge" and "task knowledge" during a pre-training phase and store this "knowledge" in their parameters (Pan and Yang 2010; Ruder 2019).[1] During a subsequent fine-tuning phase, they can then build upon this "prior knowledge" to learn new tasks with less data. Put differently, a model's parameters can represent statistical patterns of word probabilities ("language knowledge"), link word correlations to specific classes ("task knowledge") and later reuse these parameter representations for new tasks ("knowledge transfer").

In the political science literature, the use of shallow "language knowledge" through pre-trained word embeddings has become increasingly popular (Rodman 2020; Rodriguez and Spirling 2022), whereas the investigation of deep "language knowledge" and models like BERT (Bidirectional Encoder Representations from Transformers) has only started very recently on selected tasks (Bestvater and Monroe 2022; Licht 2023; Widmann and Wich 2022). We are not aware of political science literature on "task knowledge."

This paper therefore makes the following contributions. We systematically analyze: the benefits of transfer learning across a wide range of tasks and datasets relevant for political scientists; the importance of "task knowledge" as a second component of transfer learning; the impact of transfer learning on imbalanced data; and how much training data, and therefore annotation labor, different algorithms require. Our insights can help future research projects estimate their data requirements with different methods.

To test the theoretical advantages of transfer learning, we systematically compare the performance of two classical supervised algorithms (support vector machine [SVM] and logistic regression) to two transfer learning models (BERT-base and BERT-NLI) on eight tasks from five widely used political science datasets.

Our analysis empirically demonstrates the benefits of transfer learning. BERT-NLI outperforms classical models by 10.7 to 18.3 percentage points (F1 Macro) on average when 100 to 2,500 annotated data points are available. BERT-NLI achieves similar average F1 Macro performance with 500 data points as classical models with around 5,000 data points. We also show that BERT-NLI performs better with very little training data ($\leq$1,000), while BERT-base is better when more data are available. Moreover, we find that "shallow knowledge transfer" through word embeddings also improves classical models. Lastly, we show that transfer learning is particularly beneficial for imbalanced data. These benefits of transfer learning robustly apply across a wide range of datasets and tasks.

We conclude by discussing limitations of deep transfer learning and by outlining new opportunities for political science research. To simplify the reuse of BERT-NLI in future research projects, we open-source our code[2], general purpose BERT-NLI models[3] and provide advice for future research projects.

2. Supervised Machine Learning from a Transfer Learning Perspective

2.1. Supervised Machine Learning in Political Science

The rich text-as-data literature demonstrates the wide variety of methods in the toolkit of political scientists: supervised or unsupervised ideological scaling; exploratory text classification with

[1] Note that we only use the word "knowledge" to help create an intuitive understanding of transfer learning without too much jargon. Language models (i.e., pre-trained algorithms) do not "know" or "understand" anything in a deeper sense. The machine-learning process is essentially a sequence of parameter updates to optimize the statistical solution of a very specific task. Some authors colloquially call this internal parameter representation "knowledge." For a more formal discussion of transfer learning, see Pan and Yang (2010) and Ruder (2019).

[2] An easy-to-use Jupyter notebook for training your own BERT-NLI model and the full reproduction code is available at https://github.com/MoritzLaurer/less-annotating-with-bert-nli.

[3] Several models are available at https://huggingface.co/MoritzLaurer.

https://doi.org/10.1017/pan.2023.20 Published online by Cambridge University Press

unsupervised machine learning; or text classification with prior categories with dictionaries or supervised machine learning (Benoit 2020; Chatsiou and Mikhaylov 2020; Grimmer and Stewart 2013; Lucas *et al*. 2015; Van Atteveldt *et al*. 2022; Wilkerson and Casas 2017). This paper focuses on one specific group of approaches: text classification with prior categories with supervised machine learning.

In the social sciences, supervised machine-learning projects normally start with a substantive research question which requires the repetition of a specific classification task on a large textual corpus. Researchers might want to: explain Russian foreign policy by classifying thousands of statements from military and political elites into "activist" versus "conservative" positions (Stewart and Zhukov 2009); or understand delegation of power in the EU and classify legal provisions into categories of delegation (Anastasopoulos and Bertelli 2020); or predict election results and need to classify thousands of tweets into sentiment categories to approximate twitter users' preferences toward key political candidates (Ceron *et al*. 2014). These research projects required the classification of thousands of texts in topical, sentiment, or other conceptual categories (classes) tailored to a specific substantive research interest.

Using supervised machine learning to support this process roughly involves the following steps: A tailored classification task is developed, for example, through iterative discussions resulting in a codebook; experts or crowd workers implement the classification task by manually annotating a smaller set of texts (training and test data); a supervised machine-learning model is trained and tested on this manually annotated data to reproduce the human annotation task; if the model's output obtains a desired level of accuracy and validity, it can be used to automatically reproduce the task on very large unseen text corpora. If implemented well, the aggregate statistics created through this automatic annotation can then help answer the substantive research question.

Political scientists have mostly used a set of *classical supervised algorithms* for this process, such as SVMs, logistic regression, naïve Bayes, etc. (Benoit 2020). These classical algorithms are computationally efficient and obtain good performance if large amounts of annotated data are available (Terechshenko *et al*. 2020). Their input is usually a document-feature matrix which provides the weighted count of pre-processed words (features) per document in the training corpus. Solely based on this input, these models try to learn which feature (word) combinations are most strongly linked to a specific class (e.g., the topic "economy"). Several studies have shown the added value of these algorithms (e.g., Colleoni, Rozza, and Arvidsson 2014; Osnabrügge *et al*. 2021; Peterson and Spirling 2018).

The key disadvantage of these classical algorithms is that they start the training process without any prior "knowledge" of language or tasks. Humans know that the words "attack" and "invasion" express similar meanings, or that the words "happy" and "not happy" tend to appear in different contexts. Humans also quickly understand the task "classify this text into the category 'positive' or 'negative."' Classical models on the other hand need to learn these language patterns and tasks from scratch with the training data as the only source of information. Before training, the SVM is only an equation that can draw lines into space. A SVM has no prior internal representation of the semantic distance between the words "attack," "war," and "tree." This lack of prior "knowledge" of language and tasks is the main reason why classical supervised machine learning requires large amounts of training data.

A first solution to the "language knowledge" limitation compatible with classical algorithms was popularized in 2013 with word embeddings (Mikolov *et al*. 2013). Word embeddings represent words that are often mentioned in similar contexts with similar vectors – a proxy for semantic similarity. These embeddings can for example be used as input features for classifiers to provide them with a form of "language knowledge" and have gained popularity in political science (Rodman 2020; Rodriguez and Spirling 2022). Word embeddings alone provide, however, only "shallow language knowledge": first, the information they capture is limited. The vector

https://doi.org/10.1017/pan.2023.20 Published online by Cambridge University Press

of the word “capital” is the same, whether it appears next to the word “city,” “investment,” or “punishment.” Second, the improvement, which word embeddings offer for classical algorithms is only a different input layer: word embeddings instead of, for example, TF-IDF as input. Newer models integrate word embeddings into stacked layers of many additional vectors (parameters). These multi-layered, “deeper” architectures are designed to store more “knowledge.”

2.2. Deep Transfer Learning

Deep transfer learning tries to create “prior knowledge” by splitting the training procedure in roughly two phases: pre-training and fine-tuning (Howard and Ruder 2018). First, an algorithm is pre-trained to learn some general purpose statistical “knowledge” of language patterns in a wide variety of domains (e.g., news, books, and blogs), creating a language model. Second, this pre-trained model is fine-tuned on annotated data to learn a very specific task.[4] Transfer learning therefore has two important components (Pan and Yang 2010; Ruder 2019): (1) learning statistical patterns of language (*language representations*) and (2) learning a relevant task (*task representations*). Both types of representations are stored in the parameters of the model.

For learning general purpose *language representations*, the most prominent solution is BERT (Devlin *et al*. 2019) which is a type of transformer model (Vaswani *et al*. 2017). Transformers like BERT are first pre-trained using a very simple task such as masked language modeling (MLM), which does not require manual annotation. During MLM, some words are randomly hidden from the model and it is tasked with predicting the correct hidden words. The overall objective of this procedure is for the model’s parameters to learn statistical patterns of language (language representations) such as semantic similarities of words or context-dependent ambiguities from a wide variety of texts (see Appendix B1 of the Supplementary Material for details).

While sizeable performance increases with BERT-base models are possible based on its “language knowledge” (Devlin *et al*. 2019), data requirements are still relatively high. Widmann and Wich (2022), for example, show strong performance gains for an emotion detection task, but point out that the amount of training data is still an important limitation and that classes with less data underperform. An important reason for this is that the pre-training task BERT-base has learned (MLM) is very dissimilar to the actual final classification tasks researchers are interested in. This is why the last, task-specific layer of BERT (the task head tuned for MLM) is normally deleted entirely and reinitialized randomly before fine-tuning – which constitutes an important loss of “task knowledge” (see Appendix B of the Supplementary Material for details on BERT’s layered structure). BERT then needs to be fine-tuned on manually annotated data, to learn a new, useful task and each of its classes from scratch.

2.3. BERT-NLI – Leveraging the Full Potential of Deep Transfer Learning

More recently, methods have been proposed which do not only use prior “language knowledge,” but also prior “task knowledge” of transformers.[5] There are several different approaches using these innovations (Brown *et al*. 2020; Schick and Schütze 2021). This paper uses one approach, based on natural language inference (NLI), first proposed by Yin, Hay, and Roth (2019) and later refined, for example, by Wang *et al*. (2021)).

What is NLI? NLI is a task and data format, which consists of two input texts and three output classes. The input texts are a “context” and a “hypothesis.” The task is to determine if the hypoth-

[4] This describes the focus of the main steps. In practice, pre-training also involves learning (less relevant) task(s) and fine-tuning also involves learning the language of specific domain(s) (e.g., legal or social media texts).

[5] Note that the transfer of “task knowledge” is not inherently limited to transformers. Osnabrügge *et al*. (2021) show that the task learned by a logistic regression trained on the Manifesto Corpus can be applied to a different target corpus and that datasets with broadly useful tasks can be reused with classical models. Transfer learning is not an “either-or” category, but can be handled by different models to different extents.

https://doi.org/10.1017/pan.2023.20 Published online by Cambridge University Press

Table 1. Examples of the NLI task.

Hypothesis	Context	Class
The EU is trustworthy	The EU has betrayed its partners during the negotiations on Sunday	False
The EU is trustworthy	The US has betrayed its partners during the negotiations on Sunday	Neutral
The EU is trustworthy	Civil society praised the EU for reliably keeping its promises.	True

esis is true, false, or neutral given the context.[6] A hypothesis could be "The EU is trustworthy" with the context "The EU has betrayed its partners during the negotiations on Sunday." In this case, the correct class would be false, as the context contradicts the hypothesis. Note that it is not about finding the objective truth to a scientific hypothesis, but only about determining if the context string entails the hypothesis string (see, e.g., Table 1).

NLI has three important characteristics from a transfer learning perspective: It is data-rich, it is a universal task, and it enables label verbalization. First, NLI is a widely used and *data-rich task* in NLP. Many NLI datasets exist, and crowd-coders have created more than a million unique hypothesis-context pairs. Using this data, the pre-trained BERT-base can be further fine-tuned on the NLI classification task, creating BERT-NLI. Our BERT-NLI models are trained on a concatenation of eight general-purpose NLI datasets (around 1.2 million texts) from the NLP literature (see Appendix B3 of the Supplementary Material for details).

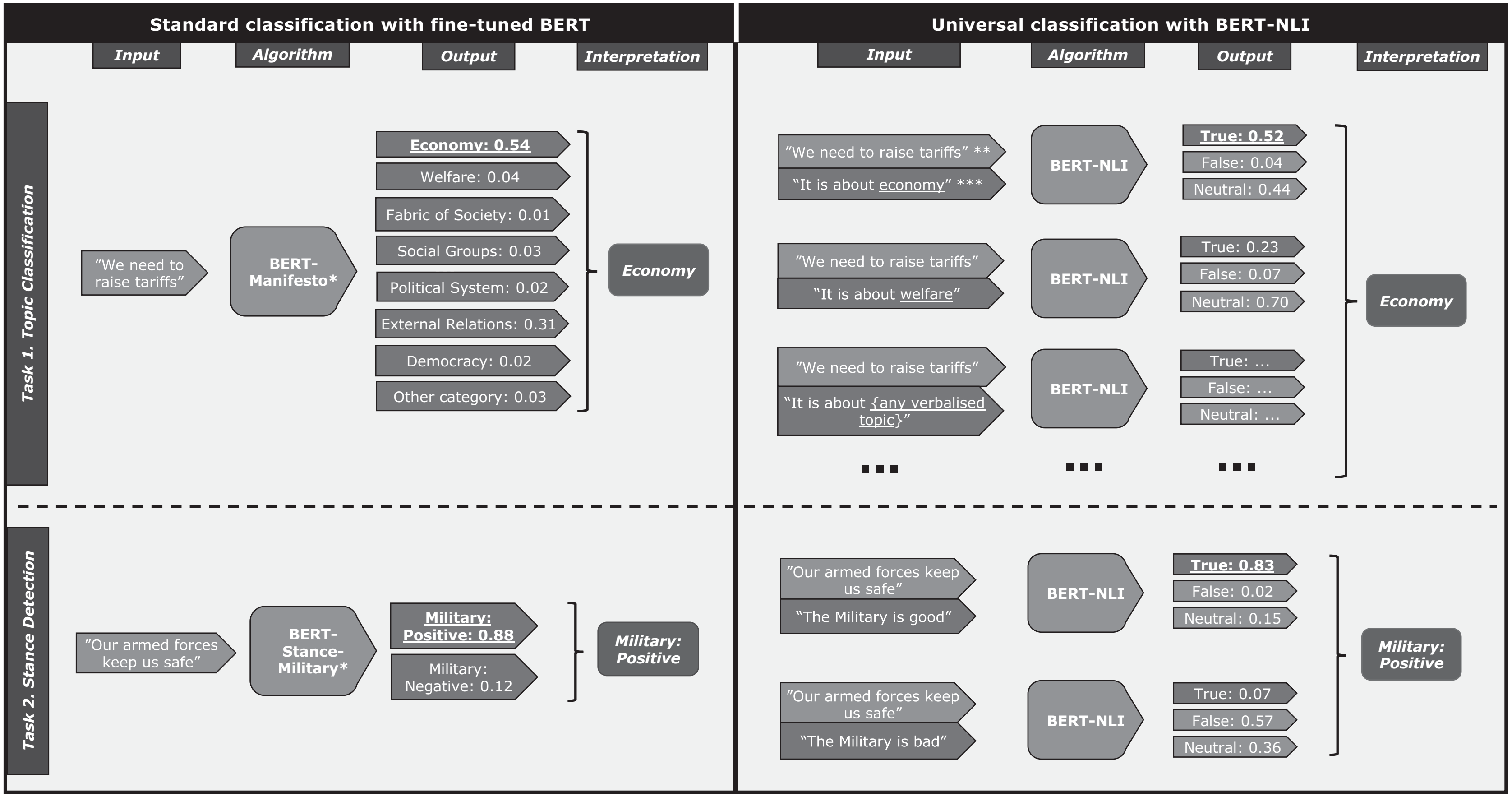
Second, NLI is a *universal task*. Almost any classification task can be converted into an NLI task. Take the text "We need to raise tariffs" and our task could be to classify this text into the eight topical classes of the Manifesto Corpus ("economy," "democracy," ...). BERT-NLI can always only execute the NLI task: predicting one of the classes true/false/neutral given a context-hypothesis pair. We can, however, translate the topic classification task into an NLI task by expressing each topical class as a "class-hypothesis," for example, "It is about economy," "It is about democracy," etc. We can then take "We need to raise tariffs" as context and test each of the class-hypotheses against this context. Each context-hypothesis pair is provided as input to BERT-NLI, which predicts the three NLI classes true/false/neutral for each class-hypothesis. We then select the topical class via the class-hypothesis that BERT-NLI predicts to be the "truest." Note that when we re-purpose BERT-NLI for other tasks like topic classification, the class-hypotheses do not have to be actually "true" in a deeper sense. The objective of reusing the classes of BERT-NLI for other tasks is only to identify the most likely downstream class relevant for the new task. The predictions for the NLI classes false and neutral class are ignored. Figure 1 illustrates how this approach enables us to solve almost *any* classification task with BERT-NLI.

Using a universal task for classification is an important advantage in situations of data scarcity. Both classical algorithms and BERT-base models need to learn the target task the researcher is interested in from scratch, with the training data as the only source of task-information. They can then only solve this very specific task. With the universal BERT-NLI classifier, almost any task can be translated into the universal NLI task format. BERT-NLI can then fully reuse the "task knowledge" it has already learned from hundreds of thousands of general-purpose NLI context-hypothesis

[6] Note that there is some variation in how the input texts and classes are called in the literature. NLI can also be called recognizing textual entailment (RTE), the "context" can be called "premise" and the three classes can be called "entailment," "contradiction," and "neutral" (Williams, Nangia, and Bowman 2018). We use the simplified vocabulary based on the instructions shown to crowd workers.

https://doi.org/10.1017/pan.2023.20 Published online by Cambridge University Press

https://doi.org/10.1017/pan.2023.20 Published online by Cambridge University Press



* BERT models fine-tuned on a specific dataset respectively. They can only predict the exact classes they have learned from their respective training data. E.g. topics from the Manifesto corpus, or stances towards the military.

** Context: Any input text treated as context for the class-hypothesis.

*** Class-hypothesis: Class verbalised by the researcher. The classes are not limited by the training data. Any task and its classes can be verbalised.

Figure 1. Illustration of standard classification versus universal NLI classification.

pairs. No task-specific parameters need to be randomly reinitialized in the task head. No "task knowledge" is lost.

This is also linked to the third important characteristic of NLI classification: *label verbalization* (Schick and Schütze 2021). Remember that human annotators always receive explicit explanations of each class in form of a codebook and can use their prior knowledge to understand the task without any examples. Standard classifiers, on the other hand, only receive examples linked to an initially meaningless number for the respective class (both classical algorithms and BERT-base). They never see the description of the classes in plain language and need to statistically guess what the underlying classification task is, only based on the training data. With the NLI task format, the class can be explicitly verbalized in the hypothesis based on the codebook (see Figure 1). More closely imitating human annotators, BERT-NLI can therefore build upon its prior language representations to understand the meaning of each class more quickly. Expressing each class in plain language provides an additional important signal to the model.

As we will show in Section 3, the combination of transformers, self-supervised pretraining, intermediate training on the data-rich NLI task, reformatting of target tasks into the universal NLI task, and label verbalization can substantially reduce the need for task-specific training data.

3. Empirical Analyses

3.1. Setup of Empirical Analyses: Data and Algorithms

To investigate the effects of transfer learning, we analyze a diverse group of datasets, representing typical classification tasks which political scientists are interested in. The datasets vary in size, domain, unit of analysis, and task-specific research interest (see Table 2). For all datasets, the overall task for human coders was to classify a text into one of multiple predefined classes of substantive political interest. Additional details on each dataset are provided in Appendix A of the Supplementary Material.

Different data pre-processing steps were tested. One objective during pre-processing is to align the classifier input more closely with the input human annotators receive. In some datasets, the unit of analysis for classification are individual quasi-sentences[7] extracted from longer speeches or party manifestos (Burst *et al*. 2020; Policy Agendas Project 2015). Human coders did, however, not interpret these quasi-sentences in isolation, but after reading the preceding (and following) text. Inspired by Bilbao-Jayo and Almeida (2018), we therefore test each algorithm with two types of inputs during hyperparameter search: only the single annotated quasi-sentence, or the quasi-sentence concatenated with its preceding and following sentence. See Appendix E of the Supplementary Material for other pre-processing steps for each algorithm.

3.1.1. Algorithms. Each dataset is analyzed with the following algorithms:

- Classical algorithms: SVM and logistic regression – two widely used algorithms to represent classical approaches. For each classical algorithm, we test two types of feature representations: TFIDF vectorization and average word embeddings (see Appendix E4 of the Supplementary Material). Word embeddings provide a shallow form of "language knowledge."[8]
- A standard transformer model: We use DeBERTaV3-base, which is an improved version of the original BERT trained on more data, with a better pre-training objective than MLM and some architectural improvements (He, Gao, and Chen 2021, see Appendix B2 of the Supplementary Material for details).

[7] A quasi-sentence is an entire sentence or a part of a sentence that represents one semantic unit. If one sentence contains two concepts of interest, it is split into two quasi-sentences.

[8] We use pre-trained GloVe embeddings (Pennington, Socher, and Manning 2014) provided by the SpaCy library (see en_core_web_lg-3.2.0, Montani *et al*. 2022), a widely used type of word embedding (Rodriguez and Spirling 2022).

https://doi.org/10.1017/pan.2023.20 Published online by Cambridge University Press

https://doi.org/10.1017/pan.2023.20 Published online by Cambridge University Press

Table 2. Key political datasets used in the analysis.

Dataset	Task	Domain	Unit of analysis	Includes context?	Avg. text length	Data points Train/Test
Manifesto Corpus (Burst *et al*. 2020)	Classify text in 8 general topics	Party manifestos	Quasi-sentences	Yes	116 characters (348 with context)	12,1570 all 88,158 train 33,412 test
Sentiment Economy News (Barberá *et al*. 2021)	Differentiate if economy is performing well or badly according to the text (2 classes)	News articles	News headline and first paragraphs	No	1,624 cha.	3,382 all 3,000 train 382 test
US State of the Union Speeches (Policy Agendas Project 2015)	Classify text in policy topics (22 classes)	Presidential speeches	Quasi-sentences	Yes	116 cha. (347 with context)	21,641 all 15,207 train 6,434 test
US Supreme Court Cases (Policy Agendas Project 2014)	Classify text in policy topics (20 classes)	Law, summaries of court cases and rulings	Court case summaries (multiple paragraphs)	No	2,456 cha.	7,752 all 5,236 train 2,326 test
CoronaNet (Cheng *et al*. 2020)	Classify text in types of policy measures against COVID-19 (20 classes)	Research assistant texts and copies from news and government sources	One or multiple sentences	No	297 cha.	48,998 all 34,298 train 14,700 test
Manifesto stances toward the military (subsets of Burst *et al*. 2020)	Identify stance toward the *simple topic* “military.” (3 classes: positive/negative/ unrelated).	Party manifestos	Quasi-sentences	Yes	Similar to Manifesto Corpus above	13,507 all 3,970 train 9,537 test
Manifesto stances toward protectionism (subsets of Burst *et al*. 2020).	Identify stance toward the *concept* “protectionism” (3 classes: positive/negative/ unrelated).	Party manifestos	Quasi-sentences	Yes	Similar to Manifesto Corpus above	5,878 all 2,116 train 3,762 test
Manifesto stances toward traditional morality (subsets of Burst *et al*. 2020).	Identify stance toward the *complex concept* “traditional morality” (3 classes: positive/negative/ unrelated).	Party manifestos	Quasi-sentences	Yes	Similar to Manifesto Corpus above	7,478 all 3,188 train 4,290 test

- An NLI-transformer: We fine-tune DeBERTaV3-base on 1.279.665 NLI hypothesis-context pairs from eight existing general-purpose NLI datasets (“BERT-NLI,” see Appendix B3 of the Supplementary Material).[9]

3.1.2. *Converting Political Science Tasks to NLI Format and Fine-Tuning BERT-NLI.* Specifically for fine-tuning BERT-NLI, the following steps were required. First, we read the codebook for each task and manually formulate one hypothesis corresponding to each class. For example, Barberá et al. (2021) asked coders to determine if a news article contains positive or negative indications on the performance of the U.S. economy. Based on the codebook, we therefore formulated the two class-hypotheses “The economy is performing well overall” and “The economy is performing badly overall.”[10] Second, we optionally write a simple script to reformat the target texts to increase the natural language fit between the class-hypothesis and the target (con)text, if necessary.[11] Third, we fine-tune the general-purpose BERT-NLI model on, for example, 500 annotated texts from the manifesto-military dataset. To this end, we match each text with the class-hypothesis, we know to be “true” based on the existing annotations and assign the label “true.” In addition, we also match each text with one random “not-true” class-hypothesis and assign the label “neutral.” This avoids that BERT-NLI learns to only predict the class “true” and provides a convenient means for data augmentation. The result is, for example, BERT-NLI-manifesto-military, which both “knows” the general NLI task and the specific manifesto-military task reformatted to NLI. Fourth, the fine-tuned model can then be applied to texts in a test set. As illustrated in Figure 1, each test text is fed into BERT-NLI exactly N times, once with each of the N different class-hypotheses. The class for which the hypothesis is the most “true” is selected.

Note that this approach allows us to further align the classifier input with the human annotator input: each human coder based their annotations on instructions in a codebook and with BERT-NLI, we can provide these coding instructions to the model via the class-hypotheses (see “label verbalization” above and Appendix B of the Supplementary Material).

3.1.3. *Comparative Analysis Pipeline and Metrics.* The objective of our analysis is to determine how much data, and therefore annotation labor, is necessary to obtain a desired level of performance on diverse classification tasks and imbalanced data. To ensure comparability and reproducibility across datasets and algorithms, each dataset is analyzed based on the same script: the random training sample size is successively increased from 0 to 10,000 texts, hyperparameters are tuned on a validation set, final performance is tested on a holdout test set. We assess uncertainty by taking three random training samples and report standard deviation (see Appendix C of the Supplementary Material).

We evaluate each model and task with multiple metrics (following the implementations by Pedregosa *et al*. 2011). Firstly, *accuracy* counts the overall fraction of correct predictions (and is equivalent to F1 Micro). The disadvantage of accuracy is that it overestimates the performance of classifiers overpredicting majority classes and neglecting minority classes. On three of our tasks, a baseline model that only predicts the majority class would already achieve above 90% accuracy due to high data imbalance. We assume that in most social science use-cases, all classes included in a task are of roughly similar importance, making accuracy a misleading metric for performance. Secondly, *balanced accuracy* calculates accuracy for each class separately and then takes the average of each per-class accuracy score (equivalent to “Recall Macro”). This gives equal weight

[9] The model is available at https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-docnli-ling-2c.

[10] In practice, we tested different hypothesis formulations during hyperparameter search (see ppendices B and E of the Supplementary Material).

[11] For some tasks, we found that reformatting the context to “The quote: ‘{context}”’ and formulating the hypotheses as “The quote is about ...” increases the natural language fit between hypothesis and context, which increases performance (see Appendix B of the Supplementary Material). The literature uses less natural formulations like “It is about ...” (Yin *et al*. 2019).

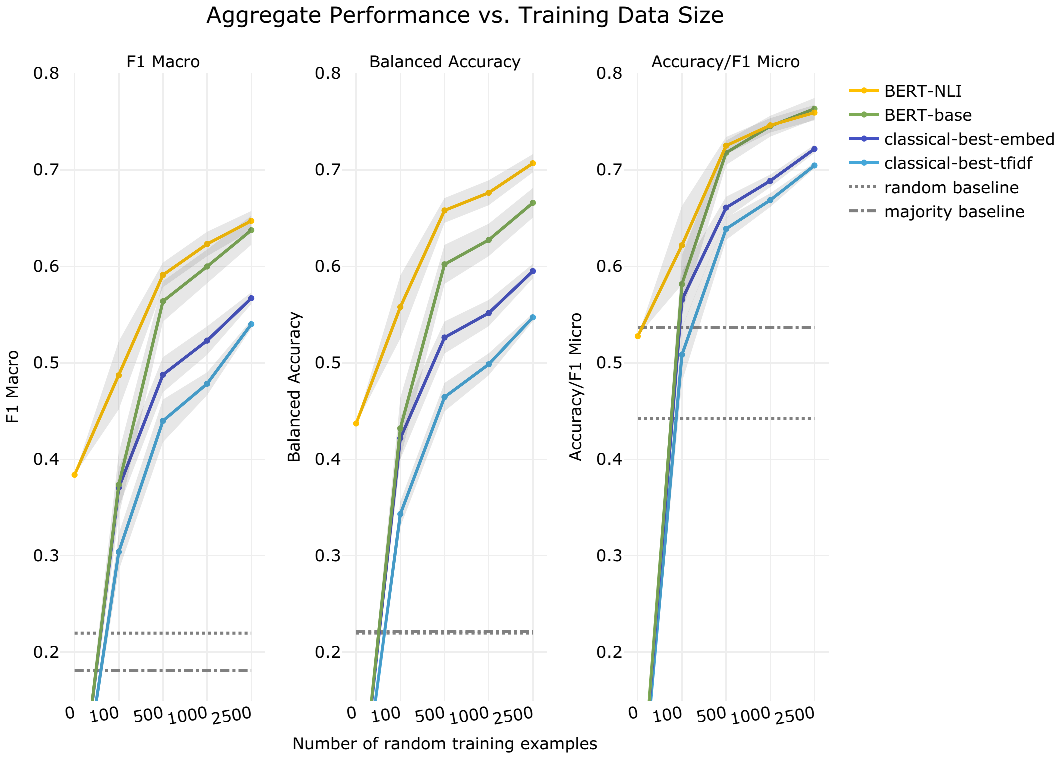


Figure 2. Average performance across eight tasks versus training data size.
The “classical-best” lines display the results from either the SVM or logistic regression, whichever is better. Note that four datasets contain more than 2,500 data points (see Figure 3).

to all classes independently of their size and is a more suitable metric, assuming that classes have similar substantive value. A characteristic of balanced accuracy is that it is higher for classifiers with less false negatives (high “Recall”) but does not properly account for false positives (risk of lower “Precision”). Balanced accuracy empirically favors classifiers that predict many minority classes well but perform less well on a few majority classes (Appendix D1 of the Supplementary Material). Thirdly, *F1 Macro* is a metric that tries to remedy this issue. It is the harmonic mean of Precision and Recall and gives equal weight to all classes independently of their size. Appendix D of the Supplementary Material provides a more detailed empirical discussion and data, including other metrics like Cohen’s Kappa. We conclude that F1 Macro is the most adequate metric for many social science use-cases of supervised machine learning and we therefore use it as the primary metric in this paper, while also reporting other metrics.[12]

3.2. Empirical Results

Figure 2 displays the aggregate average scores across all datasets. Figure 3 displays the results per dataset (see Appendix D of the Supplementary Material for detailed metrics). We focus on two main aspects across tasks: overall data efficiency and ability to handle imbalanced data.

Regarding data efficiency, deep transfer-learning models perform significantly better with less data than classical models across all tasks. The results show that BERT-NLI outperforms the classical models with TF-IDF by 10.7 to 18.3 percentage points on average (F1 Macro) when 100 to 2,500 annotated data points are available (7.9 to 12.4 with BERT-base). Classical models can be improved by leveraging shallow “language knowledge” from averaged word embeddings, but a performance difference of 8.0 to 11.7 F1 Macro remains (0.4 to 7.7 with BERT-base). The results

[12] Note that the importance of different classes might vary in different substantive research projects and researchers can make more nuanced decisions on the weight they attribute to different classes.

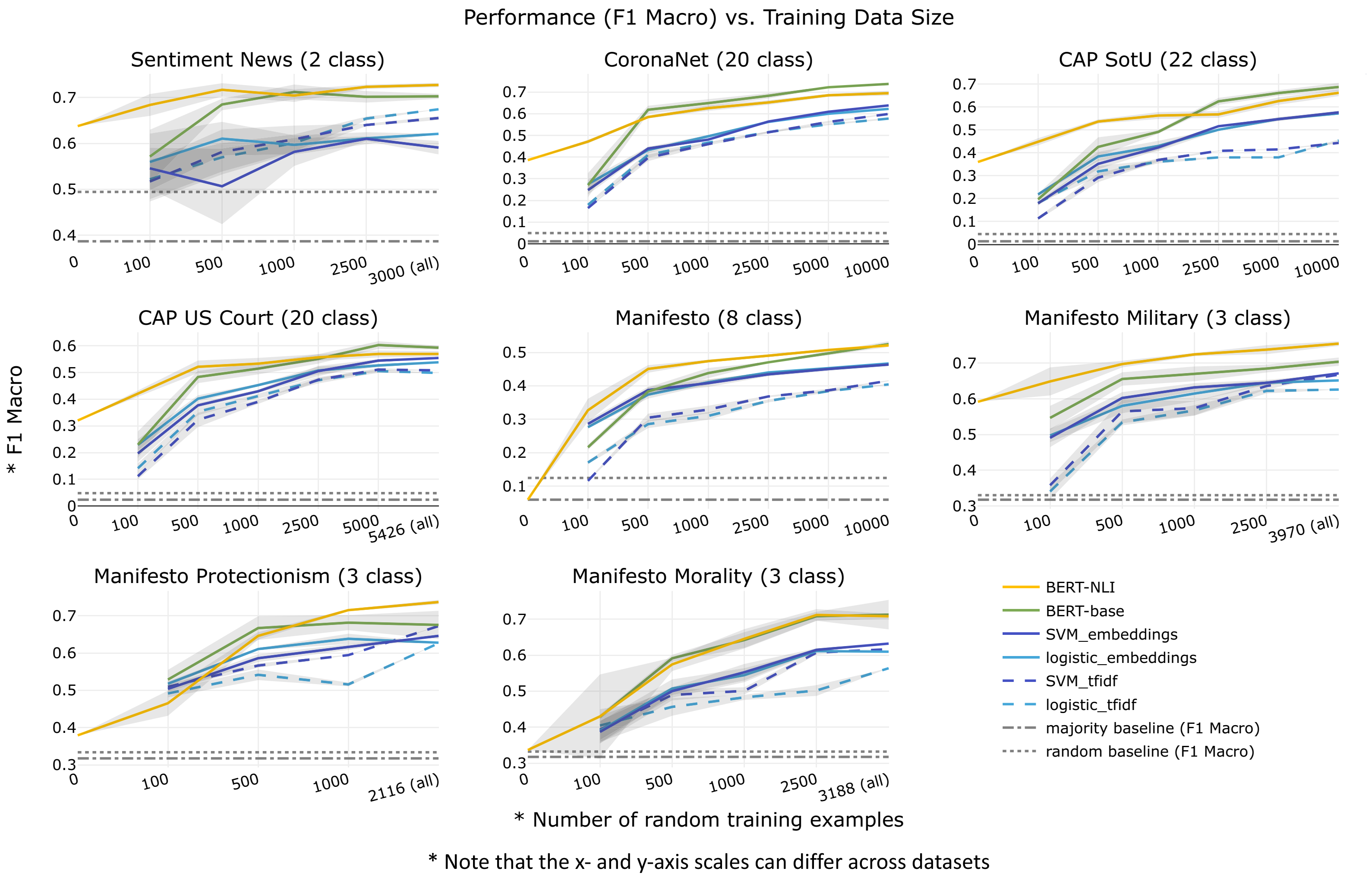


Figure 3. Performance per task versus training data size (F1 Macro).

indicate that BERT-NLI achieves similar average F1 Macro performance with 500 data points as the classical models with around 5,000 data points.[13] The performance difference remains, as larger amounts of data are sampled (5,000–10,000, see Figure 3 and Appendix D3 of the Supplementary Material) and applies across domains, units of analysis and tasks.

Moreover, the more transfer learning components a model is using, the better it becomes at handling imbalanced data. We demonstrate this by comparing accuracy/F1 Micro to F1 Macro averaged across the data intervals 100 to 2,500. Higher improvements with F1 Macro indicate an improved ability to handle imbalanced data. When "shallow language knowledge" with word embeddings is added to classical model instead of TFIDF, F1 Macro is increased by +4.6 percentage points, while accuracy/F1 Micro is only increased by +2.9 – a +1.7 higher improvement for F1 Macro. With BERT-base and its "deep language knowledge," the improvement over classical TFIDF is +7.2 with accuracy/F1 Micro and +10.3 with F1 Macro – a +3.1 higher improvement for F1 Macro. With BERT-NLI and its additional "task knowledge," the improvement is +8.3 with accuracy/F1 Micro and 14.6 with F1 Macro – a +6.3 higher improvement for F1 Macro. The higher F1 Macro score improvements compared to accuracy/F1 Micro indicates that transfer learning reduces reliance on majority classes. Good classifiers should perform similarly across all classes a researcher is interested in. Appendix D1 of the Supplementary Material provides additional data demonstrating that, when more transfer learning components are added, the performance on different classes becomes less varied.

This has two main reasons: First, both BERT variants (and word embeddings) require fewer examples for the words used in minority classes thanks to their prior representations of, for example, synonyms and semantic similarities of texts ("language knowledge"). Second, BERT-NLI performs better on F1 Macro and especially balanced accuracy and its performance across classes is least varied. Its prior "task knowledge" further reduces the need for data for smaller classes. In Appendix D1 of the Supplementary Material, we show empirically that the comparatively high performance of BERT-NLI on balanced accuracy is due to higher performance on many smaller classes compared to few majority classes. BERT-NLI can already predict a class without a single class example in the data ("zero-shot classification"). It does not need to learn each class for the new task since it uses the universal NLI task where classes are expressed in hypotheses verbalizing the codebook. This capability is also illustrated in Figures 2 and 3 by the metrics with zero training examples.

Note that our metrics are based on fully random training data samples, which do not always contain examples for all classes, especially for datasets with many classes. This simulates a typical challenge social scientists are facing, where random sampling is common and even advanced sampling techniques like active learning require an initial random sampling step (Miller *et al.* 2020). Transfer learning and especially prior "task knowledge" can therefore become another tool in our toolbox to address the issue of imbalanced data. Also note that the values for accuracy/F1 Micro are significantly higher than for F1 Macro for all models and only reporting accuracy/F1 Micro provides a misleading picture of actual performance on imbalanced data.

How to choose between BERT-base and BERT-NLI? The main criteria are the amount of training data and the degree of data imbalance. BERT-NLI is useful in situations where little and very imbalanced data is available (≤1,000). As more data becomes available to learn the new task (and minority classes) from scratch, it seems advisable to use the simpler BERT-base model given the converging performance (≥2,000). BERT-NLI has a tendency to perform better on (many) minority classes, while performing less well on (few) majority classes – which can be good or bad, depending on the use-case (see Appendix D of the Supplementary Material). Another dataset

[13] Note that the results above 2,500 data points are harder to compare, as only four datasets have enough data for the data intervals of 5,000 or more. This statement is therefore based on the performance for four datasets (see Appendix D of the Supplementary Material) as well as the overall trendline for all eight datasets.

https://doi.org/10.1017/pan.2023.20 Published online by Cambridge University Press

characteristic that can influence the value of BERT-NLI is concept complexity. BERT-NLI seems to work better when concepts are measured that can be clearly expressed in the hypotheses. For example, it performs particularly well on the manifesto-military task, measuring the stance toward the comparatively simple topic “military.” At the same time, it performs comparatively less well on manifesto-morality where the complex concept “traditional morality” is measured, which covers diverse sub-dimensions from traditional family values, religious moral values to unclear concepts like “unseemly behavior.” We assume that it is harder for BERT-NLI to map the simple language in the hypothesis to complex concepts. We discuss other factors that can influence the performance of BERT-NLI in Appendix B4 of the Supplementary Material.

Lastly, we observe that hyperparameters and text pre-processing can have an important impact on performance for all models. For example, while BERT-base models are normally trained for less than 10 epochs, we find that training for up to 100 epochs increases performance on small datasets (see Appendix E3 of the Supplementary Material for a systematic study on hyperparameters). Moreover, regarding pre-processing, if the unit of analysis are quasi-sentences, including the preceding and following sentence during pre-processing systematically increases performance for all models (Appendix E1 of the Supplementary Material); the value of word embeddings can be increased by reweighting the averaged embeddings and selecting more important words with part-of-speech tagging (Appendix E4 of the Supplementary Material); and the performance of BERT-NLI can be improved through simple pre-processing steps (Appendix B5 of the Supplementary Material).

4. Discussion of Limitations

While deep transfer learning leads to high classification performance, several limitations need to be discussed. First, deep learning models are computationally slow and require specific hardware. BERT-like transformers take several minutes to several hours to fine-tune on a high-performance GPU, while a classical model can be trained in minutes on a laptop CPU. To help alleviate this limitation, we share our experience for accessing GPUs (Appendix F of the Supplementary Material) and choosing the right hyperparameters (Appendix E3 of the Supplementary Material). Our extensive hyperparameter experiments indicate that a set of standard hyperparameters performs well across tasks and data sizes and researchers can refer to these default values to reduce computational costs.

Moreover, using BERT requires learning new software libraries. Luckily, there are relatively easy to use open-source libraries like Hugging Face transformers, which only require a moderate understanding of Python and no more than secondary education in math (Wolf *et al*. 2020).[14] Furthermore, specifically for BERT-NLI, we share our models and code. We provide several BERT-NLI models used in this paper with state-of-the-art performance on established NLI benchmarks. We invite researchers to copy and adapt our models and code to their own datasets.[15]

An additional disadvantage specifically of NLI is its reliance on human annotated NLI data, which is abundantly available in English, but less so in other languages. We also provide a multilingual BERT-NLI model pre-trained on 100 languages, but we expect it to perform less well than English-only models (Appendix B of the Supplementary Material; https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli). There are several other techniques for leveraging “prior task knowledge” which do not rely on human annotated data and could be explored in future research (Brown *et al*. 2020; Schick and Schütze 2021).

Lastly, model (pre-)training can introduce biases and impact the validity of outputs. There is a broad literature on bias in deep learning models (Blodgett *et al*. 2020) and this most likely extends

[14] Hugging Face also provides a beginner-friendly course: https://huggingface.co/course/chapter1/1.

[15] NLI models are available at https://huggingface.co/MoritzLaurer; an easy-to-use Jupyter notebook to train your own BERT-NLI model is available at https://github.com/MoritzLaurer/less-annotating-with-bert-nli.

https://doi.org/10.1017/pan.2023.20 Published online by Cambridge University Press

to political bias and NLI. It is possible, for example, that the hypotheses "The US is trustworthy" and "China is trustworthy" will result in different outputs for semantically equal inputs as one actor might have been mentioned more often in a negative context than others during (pre-)training. Political bias in deep learning is an important subject for future research. Moreover, the "black box" nature of deep learning models makes them harder to interpret. This becomes problematic when researchers want to understand why exactly a model has made a certain classification. There are some open-source libraries such as Captum (https://github.com/pytorch/captum) which can partly alleviate this issue by extracting the importance of specific features (words) for a classification decision to enable interpretations. More generally, whether the supervised machine-learning pipeline used for a specific new research question is internally and externally valid is an important additional assessment for substantive research projects (Baden *et al*. 2022).

5. Conclusion and Outlook

Lack of training data is a major hurdle for researchers who consider using supervised machine learning. This paper outlined how deep transfer learning can lower this barrier. Transformers like BERT can store information on statistical language patterns ("language knowledge") and they can be trained on a universal task like NLI to help them learn downstream tasks and classes more quickly ("task knowledge"). In contrast, classical models need to learn language and tasks from scratch with the training data as the only source of information for any new task.

We systematically test the effect of transfer learning on a range of eight tasks from five widely used political science datasets with varying size, domain, unit of analysis, and task-specific research interest. Across these eight tasks, BERT-NLI trained on 100 to 2,500 data points performs on average 10.7 to 18.3 percentage points better than classical models with TF-IDF vectorization (F1 Macro). We also show that leveraging the shallow "language knowledge" of averaged word embeddings with classical models improves performance compared to TF-IDF, but the difference to BERT-NLI is still large (8.0 to 11.7 F1 Macro). Our study indicates that BERT-NLI trained on 500 data points achieves similar average F1 Macro performance as classical models with around 5,000 data points. Moreover, transfer learning works particularly well for imbalanced data, as it reduces the data requirements for minority classes. We also provide advice on when to use BERT-NLI and when using a simpler BERT-base model is advisable. Researchers can use our results as a rough indicator for how much annotation labor their task could require with different methods.

Based on these empirical findings, we believe that deep transfer learning has great potential for making supervised machine learning a more valuable tool for social science research. As most research projects tackle new research questions which require new data for different tasks on mostly imbalanced data, the reduction of data requirements is a substantial benefit. Moreover, this enables researchers to spend more time on ensuring data quality rather than quantity and carefully creating test data for ensuring the validity of models. Accurate models combined with high quality datasets directly contribute to the validity of computational methods.

There are many important directions for future research this paper could not cover. This paper used random sampling for obtaining training data. Active learning can further reduce the number of required annotated examples (Miller *et al*. 2020). In fact, combinations of active learning and BERT-NLI are promising, as the zero-shot classification capabilities of BERT-NLI can be used in the first sampling round. Moreover, issues of political bias and validity need to be investigated further. Computational social scientists should become a more active part of the debate on (political) bias and validity in the machine-learning community.

Lastly, we believe that transfer learning has great potential for enabling the sharing and reusing of data and models in the computational social sciences. Datasets are traditionally mostly designed for one specific research question and fine-tuned models can hardly be reused in other research projects. Transfer learning in general and universal tasks in particular can help

https://doi.org/10.1017/pan.2023.20 Published online by Cambridge University Press

break these silos. Computational social scientists with a "transfer-learning mindset" could create general purpose datasets and models designed for a wider variety of use cases. Transfer learning opens many new venues for sharing and reuse which have yet to be explored.

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2023.20.

Data Availability Statement

All datasets used in this paper are publicly available. Replication code and cleaned data are available on GitHub (https://github.com/MoritzLaurer/less-annotating-with-bert-nli) and the code can be run interactively in a Code Ocean capsule at https://doi.org/10.24433/CO.5414009.v2 (Laurer *et al*. 2023a). A preservation copy of the same code and data can also be accessed via Dataverse at https://doi.org/10.7910/DVN/8ACDTT (Laurer *et al*. 2023b).

Acknowledgments

We would like to thank our colleagues at the Vrije Universiteit Amsterdam PolCom Group for their constructive feedback on several versions of the manuscript. Moreover, we want to thank Pablo Barbera, Camille Borrett, Slava Jankin, Hauke Licht, Drew Margolin, Cornelius Puschmann, and Mariken van der Velden for their feedback on earlier versions of the manuscript.

Funding

This work was supported by the Dutch Research Council (NWO) with a Snellius compute grant (EINF-3006) and by a Ph.D. scholarship of the Heinrich Böll Foundation.

Conflicts of Interest

The authors declare no conflicts of interest exist.

References

Anastasopoulos, L. J., and A. M. Bertelli. 2020. "Understanding Delegation through Machine Learning: A Method and Application to the European Union." *American Political Science Review* 114 (1): 291–301. https://doi.org/10.1017/S0003055419000522

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

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

Benoit, K. 2020. "Text as Data: An Overview." In *The SAGE Handbook of Research Methods in Political Science and International Relations*, edited by Luigi Curini and Robert Franzese, 461–497. London: SAGE Publications Ltd. https://doi.org/10.4135/9781526486387.n29

Bestvater, S. E., and B. L. Monroe. 2022. "Sentiment Is Not Stance: Target-Aware Opinion Classification for Political Text Analysis." *Political Analysis* 31: 1–22. https://doi.org/10.1017/pan.2022.10

Bilbao-Jayo, A., and A. Almeida. 2018. "Automatic Political Discourse Analysis with Multi-Scale Convolutional Neural Networks and Contextual Data." *International Journal of Distributed Sensor Networks* 14 (11): 155014771881182. https://doi.org/10.1177/1550147718811827

Blodgett, S. L., S. Barocas, H. Daumé III, and H. Wallach. 2020. "Language (Technology) Is Power: A Critical Survey of "Bias" in NLP." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. Doha: Association for Computational Linguistics.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, and D. Amodei. (2020). "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems*, 33, 1877–1901. Curran Associates, Inc. Retrieved from https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Burst, T., K. Werner, P. Lehmann, L. Jirka, T. Mattheiß, N. Merz, S. Regel, and L. Zehnter. 2020. "Manifesto Corpus." WZB Berlin Social Science Center. https://manifesto-project.wzb.eu/information/documents/corpus.

https://doi.org/10.1017/pan.2023.20 Published online by Cambridge University Press

Ceron, A., L. Curini, S. M. Iacus, and G. Porro. 2014. "Every Tweet Counts? How Sentiment Analysis of Social Media Can Improve Our Knowledge of Citizens' Political Preferences with an Application to Italy and France." *New Media & Society* 16 (2): 340–358. https://doi.org/10.1177/1461444813480466

Chatsiou, K., and S. Mikhaylov. (2020). "Deep learning for political science." (Vols. 1–2). *SAGE Publications Ltd*, https://doi.org/10.4135/9781526486387.

Cheng, C., J. Barceló, A. Spencer Hartnett, R. Kubinec, and L. Messerschmidt. 2020. "COVID-19 Government Response Event Dataset (CoronaNet v.1.0)." *Nature Human Behaviour* 4 (7): 756–768. https://doi.org/10.1038/s41562-020-0909-7

Colleoni, E., A. Rozza, and A. Arvidsson. 2014. "Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data: Political Homophily on Twitter." *Journal of Communication* 64 (2): 317–332. https://doi.org/10.1111/jcom.12084

Denny, M. J., and A. Spirling. 2018. "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It." *Political Analysis* 26 (2): 168–189. https://doi.org/10.1017/pan.2017.44

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Proceedings of the 2019 Conference of the North, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

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

He, P., J. Gao, and W. Chen. 2021. "DeBERTaV3: Improving DeBERTa Using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing." Preprint, arXiv:2111.09543 [Cs], December. http://arxiv.org/abs/2111.09543.

Howard, J., and S. Ruder. 2018. "Universal Language Model Fine-tuning for Text Classification." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/p18-1031.

Laurer, M., W. van Atteveldt, A. Casas, and K. Welbers, 2023a. "Replication Data for: Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI." https://doi.org/10.24433/CO.5414009.v2, Code Ocean, V2.

Laurer, M., W. van Atteveldt, A. Casas, and K. Welbers, 2023b. "Replication Data for: Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI." https://doi.org/10.7910/DVN/8ACDTT, Harvard Dataverse, V1.

Licht, H. (2023). "Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings". *Political Analysis, 1–14*. https://doi.org/10.1017/pan.2022.29.

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

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." In *Advances in Neural Information Processing Systems*, Vol. 26. Red Hook: Curran Associates, Inc. https://papers.nips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

Miller, B., F. Linder, and W. R. Mebane. 2020. "Active Learning Approaches for Labeling Text: Review and Assessment of the Performance of Active Learning Approaches." *Political Analysis* 28 (4): 532–551. https://doi.org/10.1017/pan.2020.4

Montani, I., M. Honnibal, S. Van Landeghem, A. Boyd, H. Peters, P. O'Leary McCann, et al. 2022. "Explosion/SpaCy: V3.2.4." Zenodo. https://doi.org/10.5281/ZENODO.6394862

Osnabrügge, M., E. Ash, and M. Morelli. 2021. "Cross-Domain Topic Classification for Political Texts." *Political Analysis* 31: 1–22. https://doi.org/10.1017/pan.2021.37

Pan, S. J., and Q. Yang. 2010. "A Survey on Transfer Learning." *IEEE Transactions on Knowledge and Data Engineering* 22 (10): 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (85): 2825–2830.

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

Peterson, A., and A. Spirling. 2018. "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems." *Political Analysis* 26 (1): 120–128. https://doi.org/10.1017/pan.2017.39

Policy Agendas Project. 2014. "US Supreme Court Cases." https://www.comparativeagendas.net/datasets_codebooks (accessed December 12, 2022).

Policy Agendas Project. 2015. "US State of the Union Speeches." https://www.comparativeagendas.net/datasets_codebooks (accessed December 12, 2022).

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

Rodriguez, P. L., and A. Spirling. 2022. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." *Journal of Politics* 84 (1): 101–115. https://doi.org/10.1086/715162

Ruder, S. 2019. *Neural Transfer Learning for Natural Language Processing*. Galway: National University of Ireland. https://ruder.io/thesis/neural_transfer_learning_for_nlp.pdf.

Schick, T., and H. Schütze. 2021. “Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference.” *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. https://doi.org/10.18653/v1/2021.eacl-main.20.

Stewart, B. M., and Y. M. Zhukov. 2009. “Use of Force and Civil–Military Relations in Russia: An Automated Content Analysis.” *Small Wars & Insurgencies* 20 (2): 319–343. https://doi.org/10.1080/09592310902975455

Terechshenko, Z., F. Linder, V. Padmakumar, M. Liu, J. Nagler, J. A. Tucker, and R. Bonneau. 2020. “A Comparison of Methods in Political Science Text Classification: Transfer Learning Language Models for Politics.” *SSRN Scholarly Paper ID 3724644*. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3724644

Van Atteveldt, W., D. Trilling, and C. Arcila Calderon. 2022. *Computational Analysis of Communication*, 1st ed. Hoboken, NJ: Wiley-Blackwell.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. “Attention Is All You Need.” *Advances in Neural Information Processing Systems*, 30. Curran Associates, Inc. Retrieved from https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wang, S., H. Fang, M. Khabsa, H. Mao, and H. Ma. 2021. “Entailment as Few-Shot Learner.” Preprint, arXiv:2104.14690 [Cs], April. http://arxiv.org/abs/2104.14690.

Widmann, T., and M. Wich. 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

Wilkerson, J., and A. Casas. 2017. “Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges.” *Annual Review of Political Science* 20 (1): 529–544. https://doi.org/10.1146/annurev-polisci-052615-025542

Williams, A., N. Nangia, and S. R. Bowman. 2018. “A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference.” *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers). https://doi.org/10.18653/v1/n18-1101.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, and A. Rush. (2020). “Transformers: State-of-the-Art Natural Language Processing.” Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Yin, W., J. Hay, and D. Roth. 2019. “Benchmarking Zero-Shot Text Classification: Datasets, Evaluation and Entailment Approach.” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. https://doi.org/10.18653/v1/d19-1404.

https://doi.org/10.1017/pan.2023.20 Published online by Cambridge University Press