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

07.07.22 - Manuscript

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

Department of Communication Science, Vrije Universiteit Amsterdam, Netherlands

*m.laurer@vu.nl

Abstract:

Supervised machine learning is an increasingly popular tool for analysing large political corpora. The main disadvantage of supervised machine learning is the need for thousands of manually created training data points. This issue is particularly important in the social sciences where every new research question requires the automation of a new task with new and imbalanced training data. This paper analyses how transfer learning algorithms like BERT can help address this challenge by storing information on statistical language patterns ('language knowledge'). Moreover, we show how leveraging a universal task called Natural Language Inference (NLI) further reduces data requirements ('task knowledge'). We systematically show the benefits of transfer learning on a wide range of eight tasks from five datasets. Across these eight tasks, BERT-NLI trained on 100 to 2500 data points performs on average 10.7 to 18.2 percentage points better than classical algorithms without transfer learning. Our study indicates that BERT-NLI trained on 500 data points achieves similar average performance as classical algorithms trained on around 5000 data points. 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.

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 labour. 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 (Grimmer and Stewart 2013; Lucas et al. 2015; Wilkerson and Casas 2017; Benoit 2020; Atteveldt, Trilling, and Calderon 2022).

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 an algorithm to reproduce the task on unseen text. The main challenge of this approach is the creation of training data. Supervised algorithms 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 every new research question entails 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. 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 algorithms can learn 'language knowledge' and 'task knowledge' during a pre-training phase and store this 'knowledge' in their parameters (Ruder 2019; Pan and Yang 2010).[1] During a separate fine-tuning phase, they can then build upon this 'prior knowledge' to learn new tasks with less data. Put differently, an algorithm'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 (Rodriguez and Spirling 2022; Rheault and Cochrane 2020; Rodman 2020), while the investigation of deep 'language knowledge' and models like BERT has only started very recently on selected tasks (Widmann and Wich 2022; Bestvater and Monroe 2022; Licht, Hauke forthcoming). We are not aware of political science literature on 'task knowledge'.

This paper therefore makes the following contributions: We systematically analyse the benefits of transfer learning across a wide range of tasks and datasets relevant for political scientists; we study the importance of 'task knowledge' as a second important component of transfer learning; we systematically analyse how much training data, and therefore annotation labour, different algorithms require to help projects estimate their data

[1] Note that we only use the word 'knowledge' to help create an intuitive understanding of transfer learning without too much jargon. Algorithms do not 'know' or 'understand' anything in a deeper sense. The machine learning process is essentially a sequence of parameter updates to optimise 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 Ruder (2019) and Pan and Yang (2010).

requirements with different methods; and we specifically analyse the impact of transfer learning on imbalanced data.

To test the theoretical advantages of transfer learning, we systematically compare the performance of two classical algorithms (Support Vector Machine, Logistic Regression) to two transfer learning algorithms (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 algorithms by 10.7 to 18.2 percentage points (F1-macro) on average when 100 to 2500 annotated data points are available. BERT-NLI achieves similar average F1-macro performance with 500 data points as classical algorithms with around 5000 data points. We also show that BERT-NLI performs better with very little training data (<= 1000), while BERT-base is better when more data is available. Moreover, we find that shallow knowledge transfer through word embeddings also improves classical algorithms. 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 re-use of BERT-NLI in future research projects, we open-source our code[2], general purpose BERT-NLI algorithms[3] and provide advice for future research projects.

[2] https://github.com/MoritzLaurer/less-annotating-with-bert-nli

[3] https://huggingface.co/MoritzLaurer

2. Supervised Machine Learning from a Transfer Learning Perspective

2.1 Supervised Machine Learning in Political Science

Many excellent surveys of text as data approaches in political science exist, providing a taxonomy of the most important methods (Grimmer and Stewart 2013), focussing on specific subfields of political science (Slapin and Proksch 2014; Lucas et al. 2015), or detailing the main stages of the analysis process and more recent developments (Wilkerson and Casas 2017; Benoit 2020; Atteveldt, Trilling, and Calderon 2022; Chatsiou and Mikhaylov forthcoming). 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 unsupervised machine learning; or text classification approaches with prior categories, using dictionaries or unsupervised machine learning (Grimmer and Stewart 2013). 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' vs. '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 towards 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 algorithm is trained and tested on this manually annotated data to reproduce the human annotation task; if the algorithm'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 Support Vector Machines (SVM), 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 (for example Osnabrügge, Ash, and Morelli 2021; Peterson and Spirling 2018; Burscher, Vliegenthart, and De Vreese 2015; Colleoni, Rozza, and Arvidsson 2014).

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 algorithms 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 was popularised in 2013 with word embeddings (Mikolov et al. 2013). Words that are often mentioned in similar contexts are represented with similar vectors – a proxy for semantic similarity. These embeddings can for example be used as input features for classifiers and have gained popularity in political science (Rodriguez and Spirling 2022; Rheault and Cochrane 2020; Rodman 2020). They can provide classical classification algorithms with a shallow form of ‘language knowledge’. Word embeddings have, however, two shortcomings for supervised machine learning: first, they are static. The vector of the word “capital” is the same, whether it appears next to the word “city”, “investment” or “punishment”. Second, by themselves, they are only stand-alone numeric representations of words. Newer algorithms integrated word embeddings more deeply into algorithms specifically designed for supervised machine learning (e.g. BERT).

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, blogs). Second, this pre-trained

algorithm 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 algorithm.

For learning general purpose *language representations*, the most prominent solution is BERT (Devlin et al. 2019) which is a type of Transformer algorithm (Vaswani et al. 2017). Transformers like BERT are first pre-trained using a very simple task, which does not require manual annotation (self-supervised training), for example Masked Language Modelling (MLM). For MLM, around 15% of words (or sub-word units called "tokens") are randomly hidden behind a "[MASK]" token. The algorithm is then tasked with predicting the original word behind this mask. Concretely, the Wikipedia sentence "*Corruption is a form of dishonesty (…) which is undertaken by a person (…) in order to acquire illicit benefits or abuse power (…)*" (Wikipedia 2021) could be randomly converted to *"[MASK] is a form of dishonesty (…) which is undertaken [MASK] a person (…) in order to acquire illicit benefits or [MASK] power (…)*". The algorithm is then tasked with predicting the true word behind each mask token given the context of visible words. This is repeated millions of times on texts from Wikipedia and books (16 gigabytes of text) in the original BERT algorithm and on additional data such as news articles (76GB), texts behind popular links on Reddit (38GB) and story-like texts (31GB) in newer algorithms (e.g. He, Gao, and Chen 2021, 16). The overall objective of this procedure is for the algorithm's parameters to learn statistical patterns of language

[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).

(language representations) such as semantic similarities of words or context-dependent ambiguities from a wide variety of texts.[5]

While sizeable performance increases with BERT-base models are possible based on its 'language knowledge' (Devlin et al. 2019; Terechshenko et al. 2020), 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 models like BERT-base have been pre-trained on a very generic self-supervised task like Masked Language Modelling, which is quite different from the final classification task. Most classification tasks are very dissimilar to the task of predicting hidden words. This is why the last, task-specific layer of BERT (the classification head tuned for MLM) is normally deleted entirely and reinitialised randomly before fine-tuning – which constitutes an important loss of 'task knowledge' (see appendix B for details on BERT's layered structure). BERT then needs to be fine-tuned on manually annotated data, to learn a new and 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. There are several different approaches using these innovations (Schick and Schütze 2021; Brown et al. 2020; Ma et al. 2021). This paper uses one approach, based on Natural Language Inference (NLI), first

[5] Note that there are many other pre-training tasks and procedures (Aroca-Ouellette and Rudzicz 2020).

proposed by Yin, Hay, and Roth (2019) and later refined for example by Wang et al. (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 texts are a 'context' and a 'hypothesis'. The task is to determine if the hypothesis 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 table 1 below for examples.

Table 1 - Examples of the NLI human intelligence 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

NLI has three important characteristics from a transfer learning perspective: It is data-rich, it is a universal task, and it enables label verbalisation. 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 (appendix B). Using this data, a pre-trained Transformer can

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

be further fine-tuned on the NLI classification task, creating an NLI-Transformer (e.g. BERT-NLI).

Second, NLI is a *universal task*. Almost any classification task can be converted into an NLI task. Take the text "Stocks are soaring" and our task is to classify this text into the topical classes "Economy" or "Politics". 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'. We can automatically reformat the text to 'The quote: "Stocks are soaring" ' as context and test the two class-hypotheses "The quote is about economy" and "The quote is about politics".[7] Each context-hypothesis pair is provided as input to BERT-NLI, which predicts the three NLI classes True/False/Neutral for each class-hypothesis. Note that for our purpose of classification with less training data, we loosen the requirements of the original NLI-task. For our purpose, the class-hypotheses do not have to be actually true. We are only interested in selecting the class-hypothesis which is more likely than the other hypotheses in order to select the corresponding class. The predictions for the classes False and Neutral class are ignored. See table 2 for details.

[7] We add the delimiter string 'The quote: {text}' to the context to make the hypothesis 'The quote is about …' more natural. The literature seems to use less natural formulations like 'It is about …' (Yin, Hay, and Roth 2019) which reduce performance in our experiments.

Table 2 - Universal NLI task format and label verbalisation

Original (con)text	Original class labels	Hypotheses with verbalized classes	Prediction NLI	Prediction target classes
'The quote: "Stocks are soaring" '	Economy	"The quote is about economy"	*True:* ***78%*** False: 3% Neutral: 29%	Economy
	Politics	"The quote is about politics"	True: 54% False: 10% Neutral: 36%	
	{any_other_class}	"The quote is about {any_other_class}"	True: X% False: Y% Neutral: Z%	
'We need to stop illegal migration to maintain our national way of life'	In favour of migration	"Migration is good"	True: 3% False: 57% Neutral: 40%	Against Migration
	Against migration	"Migration is bad"	*True:* ***68%*** False: 3% Neutral: 29%	

Using a universal task for classification is an important advantage in situations of data scarcity. Both classical algorithms and BERT-base algorithms need to learn the target task the researcher is interested in from scratch, with the training data as the only source of task-information. If the target task is translated into the universal NLI task, BERT-NLI can start with the relevant 'task knowledge' it has already learned from hundreds of thousands of NLI context-hypothesis pairs. No task-specific parameters need to be randomly initialised. 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 each class as an initially meaningless number (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 verbalised in the hypothesis based on the codebook (see table 2). 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 algorithm.

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 verbalisation 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 analyse 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 3). 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.

Table 3 - 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)	121570 all 88158 train 33412 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 & first paragraphs	No	1624 cha.	3382 all 3000 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)	21641 all 15207 train 6434 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	2456 cha.	7752 all 5236 train 2326 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 & government sources	One or multiple sentences	No	297 cha.	48998 all 34298 train 14700 test
Manifesto stances towards the military (subsets of Burst et al. 2020).	Identify stance towards the *simple topic* "military" (3 classes: positive/negative/ unrelated).	Party Manifestos	Quasi-sentences	Yes	Similar to Manifesto Corpus above	13507 all 3970 train 9537 test
Manifesto stances towards protectionism (subsets of Burst et al. 2020).	Identify stance towards the *concept* "protectionism" (3 classes: positive/negative/ unrelated).	Party Manifestos	Quasi-sentences	Yes	Similar to Manifesto Corpus above	5878 all 2116 train 3762 test
Manifesto stances towards traditional morality (subsets of Burst et al. 2020).	Identify stance towards the *complex concept* "traditional morality" (3	Party Manifestos	Quasi-sentences	Yes	Similar to Manifesto Corpus above	7478 all 3188 train 4290 test

	classes: positive/negative/ unrelated).					

We prepare the input for the algorithms to be as close as possible to the input human coders had received during annotation. First, for some datasets (Burst et al. 2020; Policy Agendas Project 2015), the unit of analysis for classification were individual quasi-sentences extracted from longer speeches or party manifestos. Human coders did, however, not interpret these quasi-sentences in isolation, but after reading the preceding (and following) text. We therefore test each algorithm with two types of inputs: only the single annotated quasi-sentence, or the quasi-sentence concatenated with its preceding and following sentence. Research indicates that including the quasi-sentence preceding the target sentence improves classification performance, even if the preceding sentence could belong to a different class (Bilbao-Jayo and Almeida 2018).

Second, each human coder based their annotations on instructions in a codebook. When using BERT-NLI, we can provide these explicit coding instructions to the algorithm via the class-hypotheses ('label verbalisation', see above). 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". Each text is tested against all possible class-hypotheses and the 'truest' hypothesis then determines the predicted class. By translating the codebooks into class-hypotheses, we can provide valuable additional information to the BERT-NLI algorithm, which human annotators usually receive via the codebook, but other algorithms cannot process (see appendix B for details on all hypotheses we tested on all datasets).

Algorithms and comparative analysis pipeline

Each dataset is analysed with the following algorithms:

- Classical algorithms: Support Vector Machines (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 E). Word embeddings provide a shallow form of 'language knowledge'.
- A standard Transformer: We use DeBERTaV3-base, which is an improved version of the original BERT (He, Gao, and Chen 2021).
- An NLI-Transformer: We fine-tune DeBERTaV3-base on 1.279.665 NLI hypothesis-context pairs ("BERT-NLI").

The objective of our analysis is to determine how much data, and therefore annotation labour, 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 analysed based on the same script: the random training sample size is successively increased from 0 to 10 000 data points, 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).

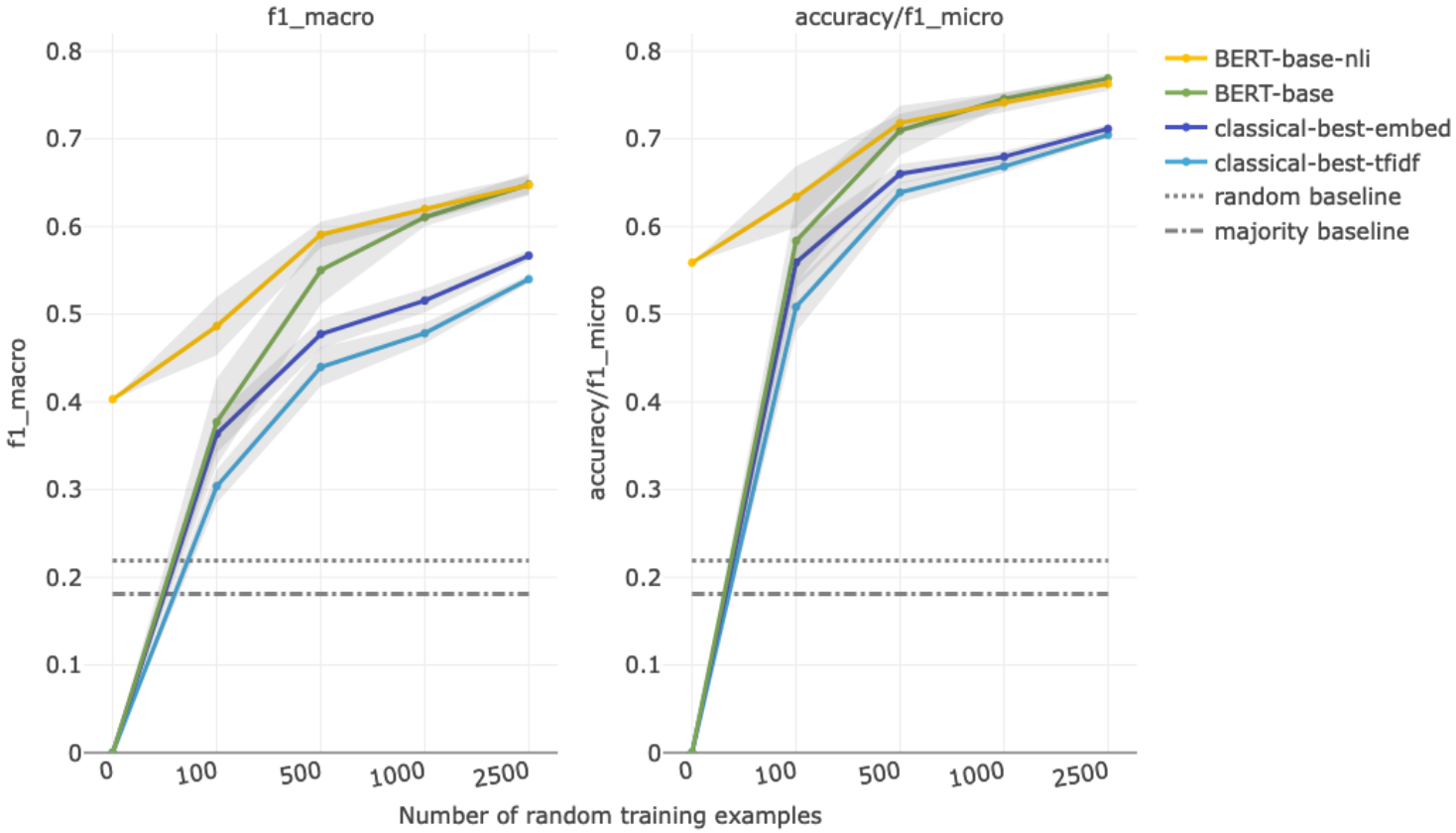
To analyse the impact of data imbalance, we analyse each algorithm with two metrics: Standard accuracy and F1-macro. Accuracy counts the fraction of correct predictions (and is equivalent to F1-micro). The disadvantage of accuracy/F1-micro is that it overestimates the performance of classifiers overpredicting majority classes and neglecting minority classes. In most social science use-cases, however, minority classes are of equivalent substantive importance as majority classes, making accuracy/F1-micro a misleading metric for

performance. We therefore use F1-macro as the primary metric. The F1 score is the harmonic mean of precision and recall. F1-macro weighs each class equally, independently of its size, and is therefore a good metric for performance on all classes with imbalanced data.

3.2 Empirical results

Figure 1 displays the aggregate average scores across all datasets. Figure 2 displays the results per dataset (see appendix D for detailed metrics). We focus on two main aspects across tasks: overall data efficiency and ability to handle imbalanced data.

Figure 1 - Average performance on eight tasks vs. training data size per algorithm



The 'classical-best' lines display the results from either the SVM or Logistic Regression, whichever is better. Note that four datasets contain more than 2500 data points, see figure 2.

Regarding data efficiency, deep transfer learning algorithms perform significantly better than classical algorithms across all tasks. The results show that BERT-NLI outperforms the classical algorithms with TFIDF by 10.7 to 18.2 percentage points on average (F1-macro) when 100 to 2500 annotated data points are available (7.3 to 13.2 with BERT-base). Classical algorithms can be improved by leveraging shallow 'language knowledge' from averaged word embeddings, but a performance difference of 8.0 to 12.3 F1-macro remains (1.4 to 9.4 with BERT-base). The results indicate that BERT-NLI achieves similar average F1-macro performance with 500 data points as the classical algorithms with around 5000 data points.[8] The performance difference remains, as larger amounts of data are sampled (5000 – 10 000, see figure 2 and appendix D) and applies across domains, units of analysis and tasks.

Moreover, transfer learning is particularly effective at handling imbalanced data, as indicated by higher improvements with F1-macro than accuracy/F1-micro. With accuracy/F1-micro, BERT-base and BERT-NLI perform +7.2 and +8.4 percentage points better than the best classical algorithm with TFIDF (averaged across the data intervals 100 to 2500). Measured with F1-macro, BERT-base and BERT-NLI perform +10.6 and +14.6 percentage points better. For classical algorithms, the shallow 'knowledge transfer' through averaged word embeddings also leads to a higher F1-macro improvement (+4.0) than accuracy/F1-micro improvement (+2.3) compared to TFIDF.

The higher F1-macro score improvements compared to accuracy/F1-micro indicates that transfer learning reduces reliance on majority classes. First, both BERT variants (and word embeddings) require fewer examples for the words used in minority classes thanks to their prior representations of e.g. synonyms and semantic similarities of texts ('language

[8] Note that the results above 2500 data points are harder to compare, as only 4 datasets have more than 2500 data points. This statement is therefore based on the performance for 4 datasets (see appendix D) as well as the overall trendline for all 8 datasets.

knowledge'). Second, BERT-NLI performs particularly well on F1-macro. Its prior 'task knowledge' further reduces the need for data for minority classes. In fact, 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, but it uses the universal NLI task where classes are expressed in hypotheses verbalising the codebook. This capability is also illustrated in figures 1 and 2 by the metrics with zero training examples.

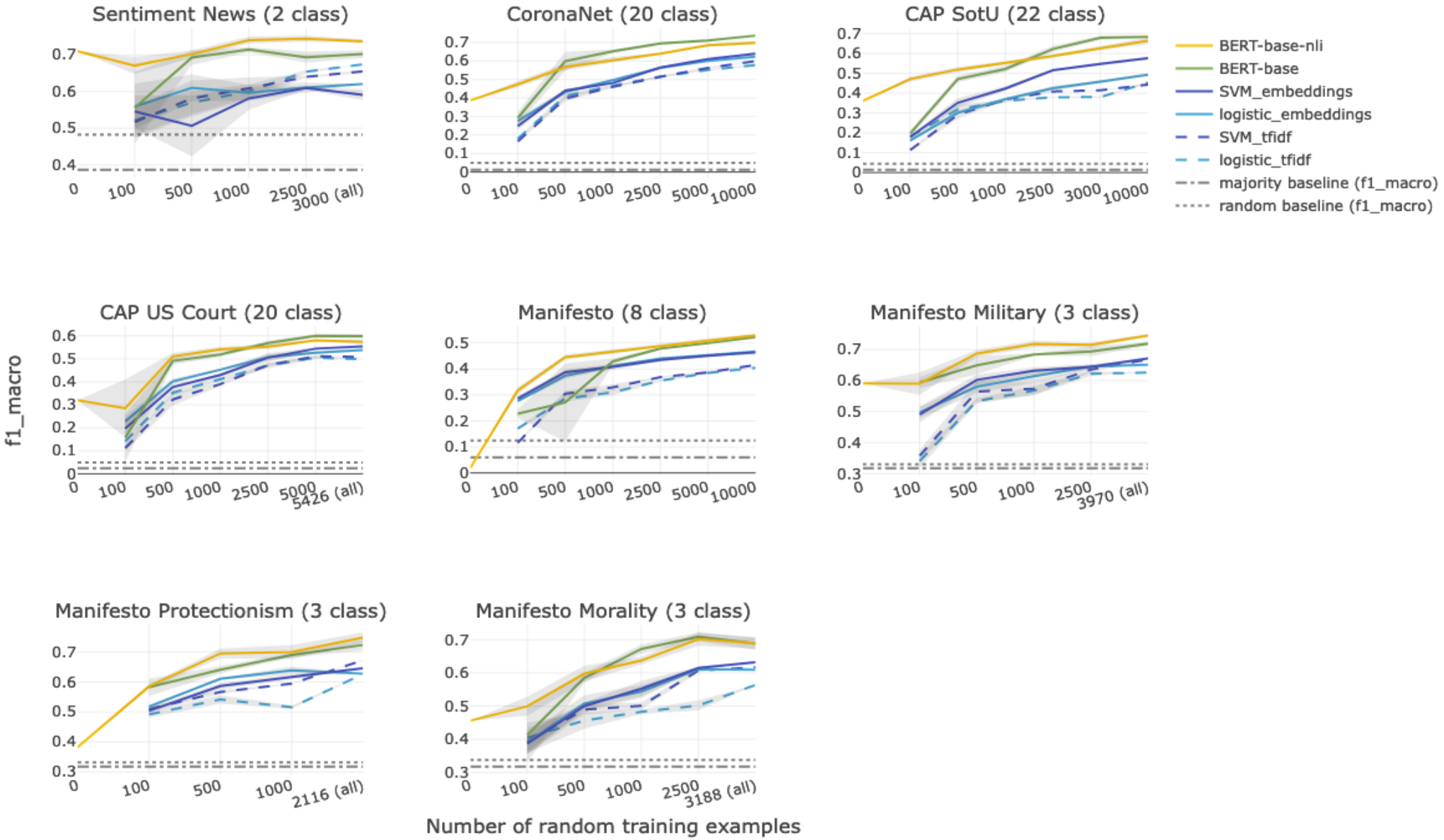
Note that our metrics are based on fully random training data samples, which do not always contain examples for all classes, especially for datasets with many classes. This simulates a typical challenge social scientists are facing, where random sampling is common and even advanced sampling techniques like active learning require an initial random sampling step (Miller, Linder, and Mebane 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 accuracy/F1-micro is significantly higher than F1-macro for all algorithms and only reporting one metric provides a misleading picture of actual performance on imbalanced data.

Overall, the comparison between the two BERT algorithms shows that BERT-NLI is useful in situations where little and imbalanced data is available (<= 1000). As more data becomes available to learn the new task (and minority classes) from scratch, the value of the universal task format decreases. At around 1000 to 2500 data points, enough data seems to be available for BERT-base to learn the new task from scratch and reusing the universal task does not add value anymore or hurts performance.[9]

[9] The main reason for this is probably that the NLI classification head always only has parameters for three general classes (True/False/Neutral), while fine-tuning BERT-base entails adding tailored parameters for each new class in the head. As more data is added, the new classification head of BERT-base can be better tailored to the new classes and therefore performs better than the generic 3-class head of BERT-NLI.

In addition, we observe that hyperparameters and text pre-processing can have an important impact on performance. For example, while BERT algorithms are normally trained for less than 10 epochs, we find that training for up to 100 epochs increases performance on small datasets (see appendix E for a systematic study on hyperparameters). Moreover, if the unit of analysis are quasi sentences, including the preceding and following sentence during pre-processing systematically increases performance.

Figure 2 - Classification performance (F1-macro) as a function of training data size per dataset



4. Discussion of Limitations

While 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 algorithm can be trained in minutes on a laptop CPU. To help alleviate this limitation, we share our experience for accessing GPUs (appendix F) and choosing the right hyperparameters (appendix E). 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).[10] Furthermore, specifically for BERT-NLI, we share our algorithms 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.[11]

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 the English models (appendix B).[12] There are several other techniques for leveraging

[10] The main library used in this paper is Hugging Face Transformers, which also provides a beginner-friendly introduction: https://huggingface.co/course/chapter1/1

[11] All our NLI models are available at https://huggingface.co/MoritzLaurer. Our code is available at https://github.com/MoritzLaurer/less-annotating-with-bert-nli.

[12] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli

‘prior task knowledge’ which do not rely on human annotated data and could be explored in future research (Schick and Schütze 2021; Brown et al. 2020; Ma et al. 2021).

Lastly, algorithm (pre-)training can introduce biases and impact the validity of outputs. There is a broad literature on bias in deep learning algorithms (Blodgett et al. 2020) and this most likely extends 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. Similarly, 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. 2021).

5. Conclusion and outlook

Lack of training data is a major hurdle for researchers considering 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 algorithms 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 2500 data points perform on average 10.7 to 18.2 percentage points better than classical algorithms with TFIDF vectorization (F1-macro). We also show that leveraging the shallow 'language knowledge' of averaged word embeddings with classical algorithms improves performance compared to TFIDF, but the difference to BERT-NLI is still large (8.0 to 12.3 F1-macro). Our study indicates that BERT-NLI trained on 500 data points achieves similar average F1-macro performance as classical algorithms with around 5000 data points. Moreover, transfer learning works particularly well for imbalanced data, as it reduces the data requirements for minority classes. Researchers can use our results as a rough indicator for how much annotation labour 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 algorithms. Accurate algorithms 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, Linder, and Mebane 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 algorithms in the computational social sciences. Datasets are traditionally only useful for one specific research question and fine-tuned algorithms can hardly be reused in other research projects. Transfer learning in general and universal tasks in particular can help break these silos. Computational social scientists with a 'transfer learning mindset' could create general purpose datasets and algorithms designed for a wider variety of use cases. Transfer learning opens many new venues for sharing and reuse which have yet to be explored.

Funding: This work was supported by the Dutch Research Council (NWO) with a the Snellius compute grant (EINF-3006).

Acknowledgements: We would like to thank our colleagues at the VU 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.

Data Availability Statement

All datasets used in this paper are publicly available. Replication code and cleaned data is available at https://github.com/MoritzLaurer/less-annotating-with-bert-nli

Conflicts of Interest: The authors are not aware of conflicts of interest.

Figure legend

Figure 2 - Average performance on 8 tasks vs. training data size

Figure 2 - Classification performance (F1-macro) as a function of training data size per dataset

Table 1 - Examples of the NLI human intelligence task

Table 2 - Universal NLI task format and label verbalisation

Table 3 - Key political datasets used in analysis

Bibliography

Anastasopoulos, L. Jason, and Anthony M. Bertelli. 2020. 'Understanding Delegation Through Machine Learning: A Method and Application to the European Union'. *American Political Science Review* 114 (1). Cambridge University Press: 291–301. doi:10.1017/S0003055419000522.

Aroca-Ouellette, Stephane, and Frank Rudzicz. 2020. 'On Losses for Modern Language Models'. *ArXiv:2010.01694 [Cs]*, October. http://arxiv.org/abs/2010.01694.

Atteveldt, Wouter van, Damian Trilling, and Carlos Arcila Calderon. 2022. *Computational Analysis of Communication*. 1st edition. Hoboken, NJ: Wiley-Blackwell.

Baden, Christian, Christian Pipal, Martijn Schoonvelde, and Mariken A. C. G van der Velden. 2021. 'Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda'. *Communication Methods and Measures* 0 (0). Routledge: 1–18. doi:10.1080/19312458.2021.2015574.

Barberá, Pablo, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. 'Automated Text Classification of News Articles: A Practical Guide'. *Political Analysis* 29 (1). Cambridge University Press: 19–42. doi:10.1017/pan.2020.8.

Benoit, Ken. 2020. 'Text as Data: An Overview'. In *The SAGE Handbook of Research Methods in Political Science and International Relations*, by Luigi Curini and Robert Franzese, 461–97. 1 Oliver's Yard, 55 City Road London EC1Y 1SP: SAGE Publications Ltd. doi:10.4135/9781526486387.n29.

Bestvater, Samuel E., and Burt L. Monroe. 2022. 'Sentiment Is Not Stance: Target-Aware Opinion Classification for Political Text Analysis'. *Political Analysis*, April. Cambridge University Press, 1–22. doi:10.1017/pan.2022.10.

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

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. 'Language (Technology) Is Power: A Critical Survey of "Bias" in NLP'. *ArXiv:2005.14050 [Cs]*, May. http://arxiv.org/abs/2005.14050.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. 'Language Models Are Few-Shot Learners'. *ArXiv:2005.14165 [Cs]*, July. http://arxiv.org/abs/2005.14165.

Burscher, Bjorn, Rens Vliegenthart, and Claes H. De Vreese. 2015. 'Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts?' *The ANNALS of the American Academy of Political and Social Science* 659 (1). SAGE Publications Inc: 122–31. doi:10.1177/0002716215569441.

Burst, Tobias, Krause Werner, Pola Lehmann, Lewandowski Jirka, Theres Mattheiß, Nicolas Merz, Sven Regel, and Lisa Zehnter. 2020. 'Manifesto Corpus'. WZB Berlin Social Science Center. https://manifesto-project.wzb.eu/information/documents/corpus.

Ceron, Andrea, Luigi Curini, Stefano M Iacus, and Giuseppe 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). SAGE Publications: 340–58. doi:10.1177/1461444813480466.

Chatsiou, Kakia, and Slava Jankin Mikhaylov. forthcoming. 'Deep Learning for Political Science'. *ArXiv:2005.06540 [Cs]*. http://arxiv.org/abs/2005.06540.

Cheng, Cindy, Joan Barceló, Allison Spencer Hartnett, Robert Kubinec, and Luca Messerschmidt. 2020. 'COVID-19 Government Response Event Dataset (CoronaNet v.1.0)'. *Nature Human Behaviour* 4 (7). Nature Publishing Group: 756–68. doi:10.1038/s41562-020-0909-7.

Colleoni, Elanor, Alessandro Rozza, and Adam 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): 317–32. doi:10.1111/jcom.12084.

Denny, Matthew J., and Arthur Spirling. 2018. 'Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It'. *Political Analysis* 26 (2). Cambridge University Press: 168–89. doi:10.1017/pan.2017.44.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding'. *ArXiv:1810.04805 [Cs]*, May. http://arxiv.org/abs/1810.04805.

Grimmer, Justin, and Brandon M. Stewart. 2013. 'Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts'. *Political Analysis* 21 (3). Cambridge University Press: 267–97. doi:10.1093/pan/mps028.

He, Pengcheng, Jianfeng Gao, and Weizhu Chen. 2021. 'DeBERTaV3: Improving DeBERTa Using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing'. *ArXiv:2111.09543 [Cs]*, December. http://arxiv.org/abs/2111.09543.

Howard, Jeremy, and Sebastian Ruder. 2018. 'Universal Language Model Fine-Tuning for Text Classification'. *ArXiv:1801.06146 [Cs, Stat]*, May. http://arxiv.org/abs/1801.06146.

Licht, Hauke. forthcoming. 'Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings'. *Political Analysis*. OSF. https://osf.io/384wr/.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. 2015. 'Computer-Assisted Text Analysis for Comparative Politics'. *Political Analysis* 23 (2): 254–77. doi:10.1093/pan/mpu019.

Ma, Tingting, Jin-Ge Yao, Chin-Yew Lin, and Tiejun Zhao. 2021. 'Issues with Entailment-Based Zero-Shot Text Classification'. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 786–96. Online: Association for Computational Linguistics. doi:10.18653/v1/2021.acl-short.99.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. 'Distributed Representations of Words and Phrases and Their Compositionality'. In *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc. https://papers.nips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

Miller, Blake, Fridolin Linder, and Walter R. Mebane. 2020. 'Active Learning Approaches for Labeling Text: Review and Assessment of the Performance of Active Learning Approaches'. *Political Analysis* 28 (4). Cambridge University Press: 532–51. doi:10.1017/pan.2020.4.

Osnabrügge, Moritz, Elliott Ash, and Massimo Morelli. 2021. 'Cross-Domain Topic Classification for Political Texts'. *Political Analysis*, October. Cambridge University Press, 1–22. doi:10.1017/pan.2021.37.

Pan, Sinno Jialin, and Qiang Yang. 2010. 'A Survey on Transfer Learning'. *IEEE Transactions on Knowledge and Data Engineering* 22 (10): 1345–59. doi:10.1109/TKDE.2009.191.

Peterson, Andrew, and Arthur Spirling. 2018. 'Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems'. *Political Analysis* 26 (1). Cambridge University Press: 120–28. doi:10.1017/pan.2017.39.

Policy Agendas Project. 2014. 'US Supreme Court Cases'. https://www.comparativeagendas.net/datasets_codebooks.

———. 2015. 'US State of the Union Speeches'. https://www.comparativeagendas.net/datasets_codebooks.

Rheault, Ludovic, and Christopher Cochrane. 2020. 'Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora'. *Political Analysis* 28 (1). Cambridge University Press: 112–33. doi:10.1017/pan.2019.26.

Rodman, Emma. 2020. 'A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors'. *Political Analysis* 28 (1). Cambridge University Press: 87–111. doi:10.1017/pan.2019.23.

Rodriguez, Pedro L., and Arthur Spirling. 2022. 'Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research'. *The Journal of Politics* 84 (1). The University of Chicago Press: 101–15. doi:10.1086/715162.

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

Schick, Timo, and Hinrich Schütze. 2021. 'Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference'. *ArXiv:2001.07676 [Cs]*, January. http://arxiv.org/abs/2001.07676.

Slapin, Jonathan B., and Sven-Oliver Proksch. 2014. 'Words as Data'. In *The Oxford Handbook of Legislative Studies*, edited by Shane Martin, Thomas Saalfeld, and Kaare W. Strøm. Oxford University Press. doi:10.1093/oxfordhb/9780199653010.013.0033.

Stewart, Brandon M., and Yuri M. Zhukov. 2009. 'Use of Force and Civil–Military Relations in Russia: An Automated Content Analysis'. *Small Wars & Insurgencies* 20 (2): 319–43. doi:10.1080/09592310902975455.

Terechshenko, Zhanna, Fridolin Linder, Vishakh Padmakumar, Michael Liu, Jonathan Nagler, Joshua A. Tucker, and Richard 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. doi:10.2139/ssrn.3724644.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. 'Attention Is All You Need'. *ArXiv:1706.03762 [Cs]*, December. http://arxiv.org/abs/1706.03762.

Wang, Sinong, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. 'Entailment as Few-Shot Learner'. *ArXiv:2104.14690 [Cs]*, April. http://arxiv.org/abs/2104.14690.

Widmann, Tobias, and Maximilian Wich. 2022. 'Creating and Comparing Dictionary, Word Embedding, and Transformer-Based Models to Measure Discrete Emotions in German Political Text'. *Political Analysis*, June. Cambridge University Press, 1–16. doi:10.1017/pan.2022.15.

Wilkerson, John, and Andreu Casas. 2017. 'Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges'. *Annual Review of Political Science* 20 (1): 529–44. doi:10.1146/annurev-polisci-052615-025542.

Williams, Adina, Nikita Nangia, and Samuel R. Bowman. 2018. 'A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference'. *ArXiv:1704.05426 [Cs]*, February. http://arxiv.org/abs/1704.05426.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2020. 'HuggingFace's Transformers: State-of-the-Art Natural Language Processing'. *ArXiv:1910.03771 [Cs]*, July. http://arxiv.org/abs/1910.03771.

Yin, Wenpeng, Jamaal Hay, and Dan Roth. 2019. 'Benchmarking Zero-Shot Text Classification: Datasets, Evaluation and Entailment Approach'. *ArXiv:1909.00161 [Cs]*, August. http://arxiv.org/abs/1909.00161.