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

20.03.2023

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

Department of Communication Science, Vrije Universiteit Amsterdam, Netherlands

*m.laurer@vu.nl

Online Appendix

Appendix A: Dataset details

Overarching data pre-processing decisions

- The train-test-split for all datasets is 70% train, 30% test, except for the Sentiment Economy dataset where the split is predefined by the dataset creators. The hyperparameter search was conducted for up to 23 runs for BERT models and up to 60 runs for classical algorithms on two random 40% validation splits of the train set for each run. To ensure reproducibility and avoid seed hacking, the same random seed (42) was maintained throughout all scripts. Where multiple random seeds were necessary, the seeds were generated with a random number generator initialised with the global random seed (42).
- For datasets with quasi-sentences as the unit of analysis (Manifesto, CAP-SotU), we tested whether including preceding and following sentences improved performance. To avoid data leakage in these cases, we did not conduct the 70-30 train-test-split on the quasi-sentence level, but on the document level.
- All texts are in English language. Multilingual classification is beyond the scope of his paper and will be addressed in future work.
- Smaller cleaning steps, such as removing texts shorter than 30 characters were conducted depending on the dataset.
- For details on all pre-processing decisions, see our GitHub repository.[1]

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

Manifesto Corpus (Burst et al. 2020)

The Comparative Manifesto Project annotates party manifestos from political parties in over 50 countries since 1945.[2] We use the data from the following English-speaking countries in the corpus: New Zealand, United Kingdom, Ireland, Australia, United States, South Africa. Our analysis is based on the dataset version 2021a and was shared with us by the Manifesto Project team. We use the categories from codebook version 4 for our analysis and convert all codes from version 5 to version 4 to harmonise categories across time. We use 4 different subsets of the manifesto corpus:

1. **Manifesto-8**: Uses eight high level domain categories (including the "Other" category). This dataset constitutes a simple topical classification task in the following categories:

Table 1 - Manifesto-8 dataset label distribution

labels	train	test	all
Welfare and Quality of Life	28421	10407	38828
Economy	22878	8186	31064
Fabric of Society	9907	3868	13775
Social Groups	8416	3255	11671
Political System	7330	3444	10774
External Relations	5979	2619	8598
Freedom and Democracy	4703	1260	5963
No other category applies	524	373	897

2. Moreover, we create three more challenging subsets: **manifesto-military, manifesto-protectionism, manifesto-morality.** We created these subsets with two objectives in mind. First, these subsets represent a more complex task beyond topic identification. Each dataset consists of three classes: texts that talk positively or negatively about a specific concept or do

[2] https://manifesto-project.wzb.eu/

not talk about the concept at all. This approximates a stance detection task. For example, manifesto-military contains texts that are positive towards the military, negative towards the military, or not about the military ("Other"). Moreover, we chose these three specific subsets, as 'Military' is a relatively simple topic, 'Protectionism' is a slightly more complex concept, and 'Traditional Morality' is a complex concept which even experts would probably have a hard time defining. The choice of concepts is intended to simulate an increase in conceptual complexity.

Secondly, these datasets are particularly imbalanced. As the datasets are so imbalanced that random sampling would have resulted in essentially only "Other" class texts, these three datasets are the only artificially sampled datasets in our paper. For the test set, we sampled the "Other" class to be ten times more frequent than the two stance-related classes combined. This simulates the common situation in the social sciences where the concepts of interest are only present in a small fraction of the target dataset. For the train-set we sampled the "other" class texts to be as frequent as the two stance-related classes combined.

Table 2 - Manifesto-military dataset label distribution

labels	**train**	**test**	**all**
Other	1985	8670	10655
Military: Positive	1623	639	2262
Military: Negative	362	228	590

Table 3 - Manifesto-protectionism dataset label distribution

labels	**train**	**test**	**all**
Other	1058	3420	4478
Protectionism: Negative	564	172	736
Protectionism: Positive	494	170	664

Table 4 - Manifesto-morality dataset label distribution

labels	train	test	all
Other	1594	3900	5494
Traditional Morality: Positive	1341	317	1658
Traditional Morality: Negative	253	73	326

Sentiment Economy News (Barberá et al. 2021)

The dataset was created by (Barberá et al. 2021) and consists of headlines and the first paragraphs of news articles.[3] Crowd workers were asked to assess, whether the text contains indications of how the US economy is performing, and if so, if this indication is positive or negative. The same data as for figure 4 in (Barberá et al. 2021) was used, where texts without an indication of the performance of the US economy were excluded. The task is therefore a binary classification task, whether a news article contains a positive or negative indication of the performance of the US economy. We use the train-test split predefined by the dataset. We pre-processed the data slightly differently than (Barberá et al. 2021), for example by removing duplicates, but our results for the classical algorithms is very similar to figure 4 in (ibid.).

Table 5 - Sentiment-economy-news dataset label distribution

labels	train	test	all
negative	2016	241	2257
positive	984	141	1125

[3] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/MXKRDE

US State of the Union Speeches (Policy Agendas Project 2015)

The dataset consists of quasi-sentences from all US State of the Union Speeches from 1946 to 2020.[4] The sentences are annotated based on 22 topical categories of the Comparative Agendas Project.[5] The underlying task is therefore a topic classification task across 22 political topics (including an "Other" class). The dataset was chosen because the CAP annotation scheme is widely used in political science and political speeches are a typical text of interest for political scientists.

Table 6 - CAP State of the Union dataset label distribution

labels	train	test	all
Other	2451	1149	3600
International Affairs	2281	833	3114
Macroeconomics	2111	956	3067
Defense	2098	719	2817
Government Operations	755	340	1095
Health	687	301	988
Education	610	281	891
Social Welfare	526	213	739
Law and Crime	501	274	775
Labor	460	358	818
Foreign Trade	404	112	516
Civil Rights	367	159	526
Energy	340	120	460
Agriculture	274	94	368
Domestic Commerce	235	112	347
Technology	222	58	280
Environment	201	90	291
Housing	195	84	279
Immigration	169	66	235

[4] https://www.comparativeagendas.net/datasets_codebooks
[5] https://www.comparativeagendas.net/pages/master-codebook

Transportation	164	53	217
Public Lands	147	55	202
Culture	9	7	16

US Supreme Court Cases (Policy Agendas Project 2014)

The dataset consists of a concatenation of the summary and ruling texts of US Supreme Court cases.[6] The texts were annotated based on 20 topical categories of the Comparative Agendas Project. The underlying task is therefore a topic classification task across 20 political topics (including an "Other" class). The dataset was chosen as it contains highly specialized legal language, and the texts are on average much longer than the other datasets (2456 characters).

Table 7 - CAP US court cases dataset label distribution

labels	**train**	**test**	**all**
Law and Crime	1701	729	2430
Civil Rights	782	336	1118
Domestic Commerce	692	296	988
Labor	488	209	697
Government Operations	391	167	558
Transportation	241	103	344
Public Lands	169	73	242
Defense	120	52	172
Immigration	111	47	158
Energy	106	45	151
Macroeconomics	98	42	140
Health	95	41	136
Environment	78	34	112
Education	65	28	93
Social Welfare	64	28	92

[6] https://www.comparativeagendas.net/datasets_codebooks

Technology	64	27	91
Foreign Trade	58	25	83
Agriculture	42	18	60
Housing	32	14	46
International Affairs	29	12	41

CoronaNet (Cheng et al. 2020)

The CoronaNet Research Project[7] compiles a database on government responses to the coronavirus for over 180 countries. Each government response against COVID-19 is annotated in one of 20 classes. In addition to the annotation, research assistants copy extracts from news and government reports or provide a brief manually written summary of the measure as proof for each annotation. These text extracts are used as input for our classifier. The dataset was chosen because it contains an atypical combination of text domains and a specialised classification task linked to COVID-19. The dataset is updated on a regular basis and we are working with a bulk download from 01.24.2022.

Table 8 - CoronaNet dataset label distribution

labels	**train**	**test**	**all**
Health Resources	4528	1940	6468
Restriction and Regulation of Businesses	3906	1674	5580
Restrictions of Mass Gatherings	2697	1156	3853
Public Awareness Measures	2315	992	3307
External Border Restrictions	2168	929	3097
Restriction and Regulation of Government Services	2084	893	2977
Quarantine	1863	799	2662
Social Distancing	1841	789	2630
Closure and Regulation of Schools	1838	788	2626
Other Policy Not Listed Above	1807	774	2581

[7] https://www.coronanet-project.org/

Lockdown	1497	642	2139
Health Testing	1224	525	1749
Internal Border Restrictions	1096	469	1565
Health Monitoring	1074	460	1534
Hygiene	930	399	1329
COVID-19 Vaccines	929	398	1327
New Task Force, Bureau or Administrative Configuration	895	384	1279
Curfew	682	292	974
Declaration of Emergency	631	271	902
Anti-Disinformation Measures	293	126	419

Appendix B: Additional details on BERT, BERT-NLI and DeBERTaV3

B1: BERT-base

Pre-training: The language representations in BERT-base is created by training the algorithm with a task called Masked Language Modelling (MLM). MLM is a self-supervised task, which does not require manual annotation and can therefore be applied to raw text data. For MLM, around 15% of words (or sub-word units called "tokens") of an input text 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). [8]

Architecture: In BERT-like Transformer algorithms, the internal parameters are organised in three main layers (Devlin et al. 2019): The vocabulary, the main trunk, and the classification head. Every input text is fed through these layers successively to produce the final output – a class prediction. (1) The algorithm's *vocabulary*: For a Transformer, a (sub-)word is a list of around 768 numbers, a vector, like the well-known word embeddings. The vocabulary layer stores around 50,000 of these vectors, one for each (sub-)word (called token) in the model's

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

vocabulary. In this first vocabulary layer, a raw input text of e.g. 20 tokens is converted into their corresponding 20 vectors. Unknown words are broken into known sub-units. A Transformer might not have the word “fundamentalism” in its vocabulary, but the tokens “fundamental” and “ism”. (2) The *main trunk*, where the vector of the word “capital” is adapted depending on its surrounding words (e.g. “punishment” or “city”). Each tokens’ vector is fed through around 12 layers and multiplied with the vectors of its surrounding tokens and other parameters in each layer. (3) The *task-specific classification head*, which condenses the internal vector representations to exactly *N* numbers: The predicted probability for each of *N* classes for a specific classification task. This last task-specific layer is deleted and randomly reinitialised for each new task (loss of ‘task knowledge’) while the other two layers are maintained (storage of ‘knowledge’).

Multilingualism: Note that the vectors in all three main layers can be tuned for monolingual or multilingual tasks. The vocabulary layer can be extended to cover tokens from many languages and scripts. Popular multilingual Transformers increase the size of the vocabulary to 250,000 tokens, and pretrain it on hundreds of GB of online texts from 100 languages at the same time (Conneau et al. 2020; He, Gao, and Chen 2021). The basic architecture remains the same, only that the representations in each layer are now multilingual. These Transformers can classify texts in 100+ languages with a performance drop of several percentage points compared to monolingual Transformers (ibid.).

B2: DeBERTaV3-base

While we often refer to “BERT” in the main text for simplicity, we actually use the newer DeBERTaV3-base model for all of our experiments. DeBERTaV3 has several advantages over

the original BERT model (He, Gao, and Chen 2021). First, it is pre-trained on more data. The original BERT model is trained on 16GB of text from Wikipedia and Books, while DeBERTa is trained on 78GB of text from Wikipedia, books, web texts like blogs, story-like texts, and news. Second, DeBERTa uses 'disentangled attention', where each token (word) is not only represented as one vector, but as two vectors: one representing the word's content and one representing its position in the text. Third, version three of DeBERTa (hence DeBERTaV3) does not use the classical masked-language-modeling objective for pre-training anymore, but uses replaced-token-detection, which is more effective at creating general language representations. The combination of these innovations and some other smaller changes make DeBERTaV3 significantly better on the GLUE benchmark and other datasets compared to the original BERT or newer models like RoBERTa or ELECTRA (He, Gao, and Chen 2021). We also conducted experiments with DeBERTaV3-large, which performed even better, but is probably too large for the hardware social scientists can normally access.

B3: BERT-NLI

Our BERT-NLI model is publicly available[9] and was trained on 1 279 665 hypothesis-premise pairs from the following public NLI datasets: MultiNLI with 393k hypothesis-premise pairs (Williams, Nangia, and Bowman 2018), FEVER-NLI with 198k pairs (Nie et al. 2020), DocNLI which consists of five NLI datasets with 942k training pairs (Yin, Radev, and Xiong 2021) and the 30k linguist-guided pairs from (Parrish et al. 2021). We exclude SNLI (570k) due to quality issues with the dataset (Bowman et al. 2015) and deduplication reduced the overall amount

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

of data points. These datasets cover domains like news articles, fictional literature, government documents, telephone conversations or image captions.

We initially experimented with including and excluding different NLI datasets and we did not notice significant differences in performance, for example when the large DocNLI dataset was not included. As a rule of thumb, including multiple datasets to increase data quantity to several hundred thousand texts and to cover a wider range of domains and dataset idiosyncrasies is beneficial for learning NLI. Indiscriminately adding more data with potential quality issues like SNLI (or DocNLI) does not necessarily add value. We have open-sourced multiple NLI models based on different NLI datasets and Transformer models with different trade-offs in speed and performance.[10]

While English data is dominant, multilingual NLI data exists as well (Conneau et al. 2018). Note that a multilingually pretrained Transformer which is then fine-tuned *only* on English NLI data still obtains 79.8% average NLI accuracy on 14 other languages from Chinese to Urdu compared to 88.2% on English. 33% is the random and majority baseline and performance can be increased by including multilingual NLI data (He, Gao, and Chen 2021). We therefore also provide a multilingual NLI model.[11]

During the training process for the NLI task, the input is always a unique context-hypothesis pair, which is fed into the Transformer as one string only separated by a separator token "[SEP]". Some examples from a popular NLI dataset are: "I am a lacto-vegetarian [SEP] I enjoy eating cheese too much to abstain from dairy" (class: neutral); or "8 million in relief in the form of emergency housing [SEP] The 8 million dollars for emergency housing was still not

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

enough to solve the problem" (class: neutral); or "At 8:23, the Boston Center controller received a third transmission from American 11 [SEP] The Boston Center controller got a third transmission from American 11" (class: true); or "Met my first girlfriend that way [SEP] I didn't meet my first girlfriend until later" (class: false) (Williams, Nangia, and Bowman 2018). Note that the Transformer will not learn anything about 'truth' in a deeper sense. It will learn language patterns which make it likely for a hypothesis to be True/False/Neutral, given a context.

Note that there is a relevant literature on the mixed quality of existing NLI datasets. Widely used datasets contain artefacts such as a high correlation between negation words and the False class, or lexical overlap and the True class, which enables algorithms to solve the task without a deeper understanding of the texts (Gururangan et al. 2018). Note that the negative impact of these quality issues is less pressing for our use-case, as we do not try to optimise general reasoning abilities, but general-purpose classification (where the hypothesis does not need to be actually true).

Another disadvantage of the universal NLI task is the linearly increasing computational costs per additional class in the target task. Each context-hypothesis pair is fed through BERT-NLI separately, multiplying the required computation by the number of classes during inference. Moreover, each class-hypothesis needs to be formulated manually and different formulations can lead to changes in performance. Interestingly enough, we noticed, that training a BERT-NLI model is faster than training a BERT-base model, as less epochs (iterations over the entire training set) are required to achieve the best performance (see the appendix on training times below). This means that BERT-NLI is slower during inference, but faster during training.

Moreover, note that, while the NLI task generally includes three classes (True / False / Neutral), we actually use a classifier that only predicts two classes (True / Not-True). Our use-

case only requires the probabilities of the True class and the difference between False and Neutral is irrelevant for our purposes. We therefore merge False and Neutral data during the NLI pre-fine-tuning step into the same "Not-True" class. This has the additional benefit that binary NLI data can be added to our NLI pre-fine-tuning step. Initial tests were conducted with a three class NLI model, but no meaningful performance differences were observed.

In order to further illustrate the difference between BERT-base and BERT-NLI, we provide an illustration of the fine-tuning process for both algorithms (see figure 1 and 2 below). This illustrates the advantage of the universal task format over fine-tuning a model (be it BERT-base or a classical algorithm) on a case-specific task like the Manifesto corpus.

Figure 1 - The training process of BERT-base and architectural implications

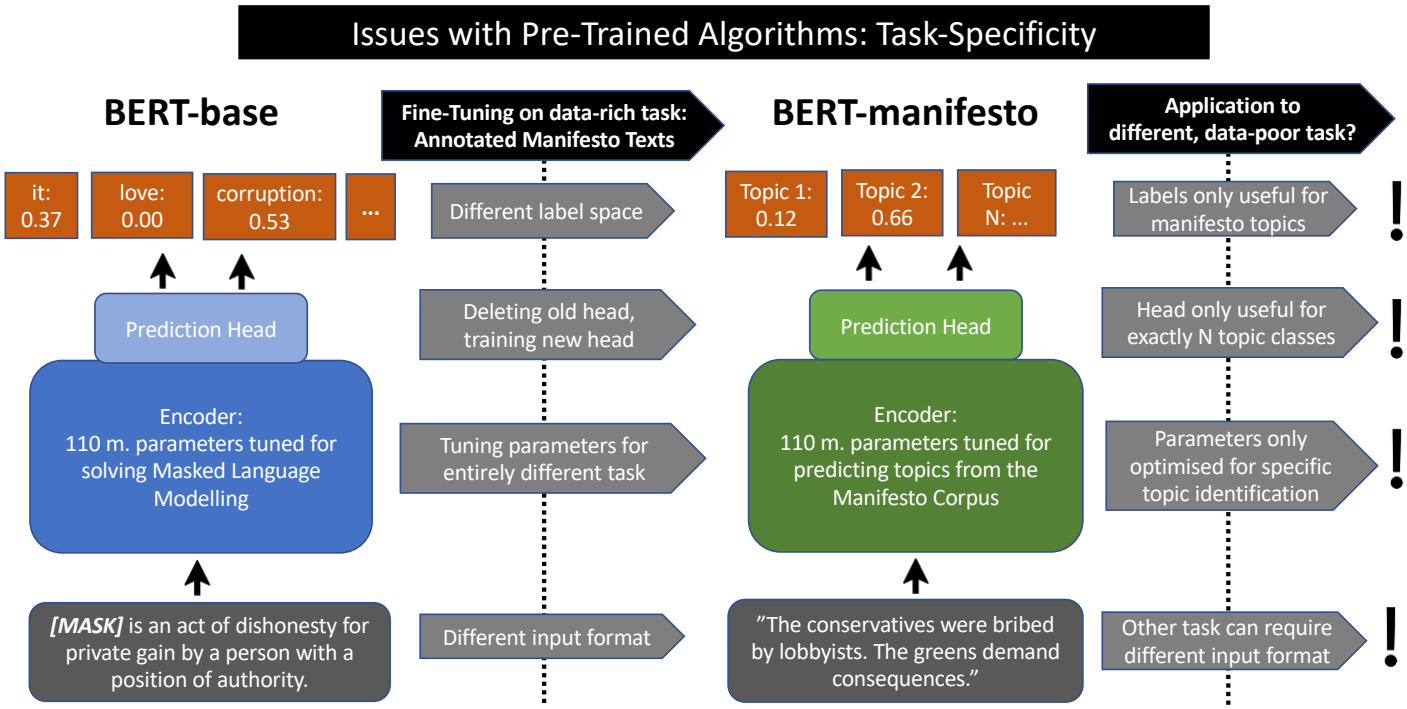
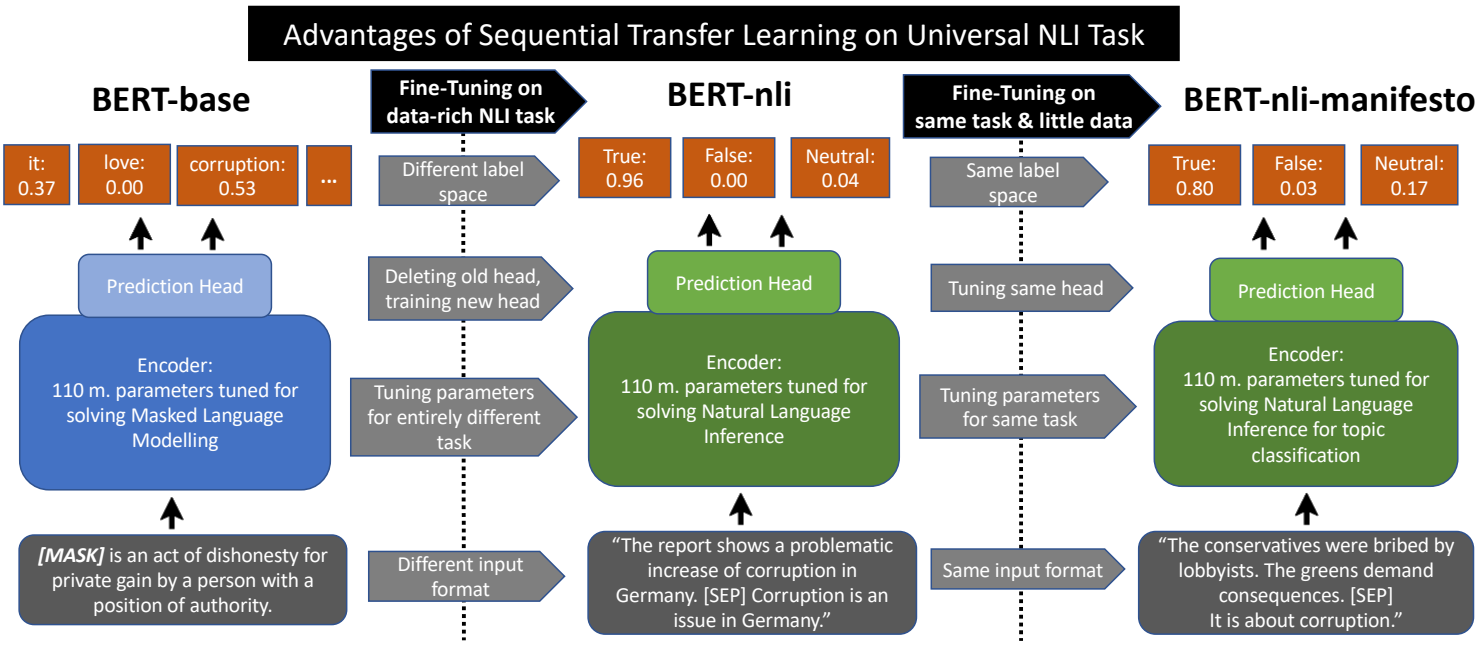


Figure 2 - The training process of BERT-NLI and architectural implications



B4: Which factors influence BERT-NLI performance?

The main factor determining the utility of BERT-NLI is the number of training data and the degree of data imbalance. BERT-NLI is better with less and very imbalanced data. It is less useful with more data and less imbalance. Besides these main factors, other factors can influence BERT-NLI's performance. We discuss these factors below. Note, however, that these assumptions are only based on eight tasks and more tasks/datasets would be necessary for firm conclusions.

(1) The complexity of the label/concept the task tries to measure. Take the three Manifesto-stance datasets. We intentionally chose three subsets from the Manifesto corpus with 3-class stance tasks, with similar text length, the same text domain, but with concepts of differing complexity. Manifesto-military measures a simple topic, 'military', that can be referenced

with relatively simple language on military equipment and disarmament or military treaties; Manifesto-protectionism measures a concept, 'protectionism'; and Manifesto-morality measures a complex concept 'traditional morality', which, based on the codebook, covers diverse sub-dimensions from traditional family values, religious moral values to unclear concepts like 'unseemly behaviour'. BERT-NLI performs best on Manifesto-military and comparatively worst on Manifesto-morality. We assume that BERT-NLI performs better on simpler concepts that can be expressed more easily in plain language (in the hypothesis) like 'military'. It is probably more difficult for the algorithm to map the language in the hypothesis to indications of complex concepts like 'traditional morality' in the target text. Similarly, BERT-NLI excels at the simple 2-class sentiment task in Sentiment News. Moreover, when comparing datasets with many classes (CoronaNet with 20, CAP, SotU with 22), BERT-NLI consistently performs worse than BERT-base with 500 texts or more on CoronaNet, while BERT-NLI performs better for one more interval on CAP SotU. We assume that this is because the CAP SotU task contains simpler categories like 'international affairs', 'defense/military', 'education' etc., while Coronanet contains more specialised classes like 'health resources' or 'social distancing', which were harder to express in clear plain language in the class-hypotheses.

(2) The specificity of the language in the target domain could play a role when very little data is available. Take the two Comparative Agendas Project datasets (CAP SotU and CAP US Court), which have the same task, but for different domains (presidential speeches and court rulings). For CAP SotU, BERT-NLI performs clearly better than BERT-base including the 1000 data point interval, while for CAP US Court, performance already becomes similar with 500 data points. This could be because court rulings have a more specialised language, and it is

harder for BERT-NLI to fruitfully map the language in the class-hypothesis to the legal language in the target texts with less training examples. Another explanation could be differences in data imbalance, but BERT-NLI still performs better compared to BERT-base on CAP SotU than CAP US Court when measured with standard accuracy/F1-micro. Another assumption could be differences in text length. CAP US Court texts are substantially longer than CAP SotU texts (2456 vs. 347 characters). At the same time, BERT-NLI performs very well on Sentiment Economy News with 1624 characters. We do assume that longer texts are more difficult for BERT-NLI, because established NLI training data tends to be only one or a few sentences long, but we cannot clearly confirm this based on our 8 tasks/datasets.

(3) Lastly, an important factor for explaining BERT-NLI performance could be the number of classes. Before starting our experiments, we had assumed that BERT-NLI's performance would decrease as the number of classes increases. We assumed that the 3-class (or 2-class) NLI task head would have difficulties handling too many classes. The findings from our 8 tasks/datasets do, however, not clearly reflect this. BERT-NLI performs better than BERT-base on CAP SotU with 22 classes (and CAP US Court, 20 classes) until including 1000 data points (F1-macro), but performs very similarly to BERT-base in almost all data intervals on Manifesto-morality with 3 classes and BERT-NLI consistently performs better than BERT-base on Manifesto-8-class. An important intervening factor seems to be the increasing data imbalance as the number of classes increases. The more classes a dataset has, the less likely it is to randomly sample enough data points for minority classes (for smaller sample size scenarios). We systematically show in appendix D that BERT-NLI performs particularly well with imbalanced data and this most likely also supports its performance for datasets with many

classes. We still believe that BERT-base's classification head is better at handling many classes, but only once enough data for minority classes is available.

Overall, however, the number of training data and the degree of data imbalance are clearly the most important factors that influences the utility of BERT-NLI. It is more useful with less and imbalance data, while we advise against using BERT-NLI if enough and more balanced data is available and the simpler BERT-base is probably preferable.

B5: Hypothesis formulation and context

Our tests showed, that BERT-NLI's performance can be increased with specific pre-processing steps. To increase the natural language fit between the target sentence and the class-hypotheses, we format the target sentence as follows: 'The quote: "{target-sentence}".'. This enables us to formulate hypotheses referring to 'The quote', such as 'The quote is about {label}'.

Moreover, this enables us to target the classifiers' attention specifically on the target sentence, in cases where we added the preceding and following sentence for additional context. The quotation mark strings provide a clear natural language delimiter for the target sentence, to distinguish it from the surrounding sentences. Note that for most classical classifiers word order does not matter, and punctuation is removed. For BERT, on the other hand, word order and punctuation are explicitly taken into account. See the table below for a concrete example.

Table 9 - Examples for pre-processing and input for BERT-NLI

Text	Class Options	Hypothesis string	Class gold
We will invest more in combating climate change. *The quote: "***We would: Ensure that adequate government funding goes to research on major environmental issues such as climate change, pollution and biodiversity loss,"** and less is spent on military research.	Other	The quote is not about military or defense	Other
	Military: Positive	The quote is positive towards the military	
	Military: Negative	The quote is negative towards the military	

Note: In the text column, bolded text represents the original target sentence which should be classified, italicised text represents delimiter strings which were added during pre-processing to focus the NLI classifier's attention on the target sentence. In this example from (Burst et al. 2020), the target sentence was classified as unrelated to the military (class 'environmental protection'), while the following sentence is 'military: negative'.

B6: NLI hypotheses tested per dataset

We formulated our hypotheses by reading the codebook of the respective dataset and verbalising the description of each class in a class-hypothesis. During initial tests, we tried several different ways of formulating hypotheses and in the end, we decided to test two formulations during hyperparameter search: a long hypothesis and a short hypothesis. The hypotheses tested during hyperparameter search for each dataset are available in the tables below. The best hypotheses based on the hyperparameter search are available in the hyperparameter tables in appendix E. In general, we noticed that shorter hypotheses worked better for smaller sample sizes, while longer hypotheses worked better for larger sample sizes.

Table 10 - Manifesto-8 hypotheses

label	hypotheses_short	hypotheses_long
Economy	The quote is about economy, or technology, or infrastructure, or free market.	The quote is about economy, free market economy, incentives, market regulation, economic planning,

		cooperation of government, employers and unions, protectionism, economic growth, technology and infrastructure, nationalisation, neoliberalism, marxism, sustainability.
External Relations	The quote is about international relations, or foreign policy, or military.	The quote is about international relations, foreign policy, anti-imperialism, military, peace, internationalism, European Union.
Fabric of Society	The quote is about law and order, or multiculturalism, or national way of life, or traditional morality.	The quote is about society, national way of life, immigration, traditional morality, law and order, civic mindedness, solidarity, multiculturalism, diversity.
Freedom and Democracy	The quote is about democracy, or freedom, or human rights, or constitutionalism.	The quote is about democracy, freedom, human rights, constitutionalism, representative or direct democracy.
Political System	The quote is about governmental efficiency, or political authority, or decentralisation, or corruption.	The quote is about political system, centralisation, governmental and administrative efficiency, political corruption, political authority.
Social Groups	The quote is about agriculture, or social groups, or labour groups, or minorities.	The quote is about social groups, labour groups, agriculture and farmers, middle class and professional groups, minority groups, women, students, old people.
Welfare and Quality of Life	The quote is about welfare, or education, or environment, or equality, or culture.	The quote is about welfare and quality of life, environmental protection, culture, equality, welfare state, education.
No other category applies	The quote is about something other than the topics economy, international relations, society, freedom and democracy, political system, social groups, welfare. It is about non of these topics.	The quote is about something other than the topics economy, international relations, society, freedom and democracy, political system, social groups, welfare. It is about non of these topics.

Table 11 - Manifesto-military hypotheses

label	hypotheses_short	hypotheses_long
Military: Positive	The quote is positive towards the military	The quote is positive towards the military, for example for military spending, defense, military treaty obligations.
Military: Negative	The quote is negative towards the military	The quote is negative towards the military, for example against military spending, for disarmament, against conscription.
Other	The quote is not about military or defense	The quote is not about military or defense

Table 12 - Manifesto-protectionism hypotheses

label	hypotheses_short	hypotheses_long
Protectionism: Positive	The quote is positive towards protectionism, for example protection of internal markets through tariffs or subsidies	The quote is positive towards protectionism, for example in favour of protection of internal markets through tariffs or export subsidies or quotas
Protectionism: Negative	The quote is negative towards protectionism, for example in favour of free trade or open markets	The quote is negative towards protectionism, for example in favour of free trade or open international markets
Other	The quote is not about protectionism or free trade	The quote is not about protectionism or free trade

Table 13 - Manifesto-morality hypotheses

label	hypotheses_short	hypotheses_long
Traditional Morality: Positive	The quote is positive towards traditional morality	The quote is positive towards traditional morality, for example in favour of traditional family values, religious institutions, or against unseemly behaviour
Traditional Morality: Negative	The quote is negative towards traditional morality	The quote is negative towards traditional morality, for example in favour of divorce or abortion, modern families, separation of church and state, modern values
Other	The quote is not about traditional morality	The quote is not about traditional morality, for example not about family values, abortion or religion

Table 14 - Sentiment-economy-news hypotheses

label	hypotheses_quote	hypotheses_complex
positive	The quote is overall positive	The economy is performing well overall
negative	The quote is overall negative	The economy is performing badly overall

Table 15 - CAP state of the union hypotheses

label	hypotheses_short	hypotheses_long
Agriculture	The quote is about agriculture.	The quote is about agriculture, for example: agricultural foreign trade, or subsidies to farmers, or food inspection and safety, or agricultural marketing, or animal and crop disease, or fisheries, or R&D.
Culture	The quote is about cultural policy.	The quote is about cultural policy.
Civil Rights	The quote is about civil rights, or minorities, or civil liberties.	The quote is about civil rights, for example: minority/gender/age/handicap discrimination, or voting rights, or freedom of speech, or privacy.
Defense	The quote is about defense, or military.	The quote is about defense, for example: defense alliances, or military intelligence, or military readiness, or nuclear arms, or military aid, or military personnel issues, or military procurement, or reserve forces, or hazardous waste, or civil defense and terrorism, or contractors, or foreign operations, or R&D.
Domestic Commerce	The quote is about banking, or finance, or commerce.	The quote is about domestic commerce, for example: banking, or securities and commodities, or consumer finance, or insurance regulation, or bankruptcy, or corporate management, or small businesses, or copyrights and patents, or disaster relief, or tourism, or consumer safety, or sports regulation, or R&D.
Education	The quote is about education.	The quote is about education, for example: higher education, or education finance, or schools, or education of underprivileged, or vocational education, or education for handicapped, or excellence, or R&D.
Energy	The quote is about energy, or electricity, or fossil fuels.	The quote is about energy, for example: nuclear energy and safety, or electricity, or natural gas & oil, or coal, or alternative and renewable energy, or conservation, or R&D.

Environment	The quote is about the environment, or water, or waste, or pollution.	The quote is about the environment, for example: drinking water, or waste disposal, or hazardous waste, or air pollution, or recycling, or species and forest protection, or conservation, or R&D.
Foreign Trade	The quote is about foreign trade.	The quote is about foreign trade, for example: trade agreements, or exports, or private investments, or competitiveness, or tariff and imports, or exchange rates.
Government Operations	The quote is about government operations, or administration.	The quote is about government operations, for example: intergovernmental relations, or agencies, or bureaucracy, or postal service, or civil employees, or appointments, or national currency, or government procurement, or government property management, or tax administration, or public scandals, or government branch relations, or political campaigns, or census, or capital city, or national holidays.
Health	The quote is about health.	The quote is about health, for example: health care reform, or health insurance, or drug industry, or medical facilities, or disease prevention, or infants and children, or mental health, or drug/alcohol/tobacco abuse, or R&D.
Housing	The quote is about community development, or housing issues.	The quote is about housing, for example: community development, or urban development, or rural housing, low-income assistance for housing, housing for veterans/elderly/homeless, or R&D.
Immigration	The quote is about migration.	The quote is about migration, for example: immigration, or refugees, or citizenship.
International Affairs	The quote is about international affairs, or foreign aid.	The quote is about international affairs, for example: foreign aid, or international resources exploitation, or developing countries, or international finance, or western Europe, or specific countries, or human rights, or international organisations, or international terrorism, or diplomats.
Labor	The quote is about employment, or labour.	The quote is about labour, for example: worker safety, or employment training, or employee benefits, or labor unions, or fair labor standards, or youth employment, or migrant and seasonal workers.
Law and Crime	The quote is about law, crime, or family issues.	The quote is about law and crime, for example: law enforcement agencies, or white collar crime, or illegal drugs, or court administration, or prisons, or juvenile crime, or child abuse, or family issues, or criminal and civil code, or police.
Macroeconomics	The quote is about macroeconomics.	The quote is about macroeconomics, for example: interest rates, or unemployment, or monetary policy, or national budget, or taxes, or industrial policy.
Other	The quote is about other, miscellaneous.	The quote is about other things, miscellaneous.
Public Lands	The quote is about public lands, or water management.	The quote is about public lands, for example: national parks, or indigenous affairs, or public lands, or water resources, or dependencies and territories.
Social Welfare	The quote is about social welfare.	The quote is about social welfare, for example: low-income assistance, or elderly assistance, or disabled assistance, or volunteer associations, or child care, or social welfare.
Technology	The quote is about space, or science, or technology, or communications.	The quote is about technology, for example: government space programs, or commercial use of space, or science transfer, or telecommunications, or regulation of media, or weather science, or computers, or internet, or R&D.
Transportation	The quote is about transportation.	The quote is about transportation, for example: mass transportation, or highways, or air travel, or railroads, or maritime, or infrastructure, or R&D.

Table 16 - CAP US court cases hypotheses

label	hypotheses_long
Agriculture	The quote is about agriculture, for example: agricultural foreign trade, or subsidies to farmers, or food inspection and safety, or agricultural marketing, or animal and crop disease, or fisheries, or R&D.
Civil Rights	The quote is about civil rights, for example: minority/gender/age/handicap discrimination, or voting rights, or freedom of speech, or privacy.
Defense	The quote is about defense, for example: defense alliances, or military intelligence, or military readiness, or nuclear arms, or military aid, or military personnel issues, or military procurement, or reserve forces, or hazardous waste, or civil defense and terrorism, or contractors, or foreign operations, or R&D.
Domestic Commerce	The quote is about domestic commerce, for example: banking, or securities and commodities, or consumer finance, or insurance regulation, or bankruptcy, or corporate management, or small businesses, or copyrights and patents, or disaster relief, or tourism, or consumer safety, or sports regulation, or R&D.
Education	The quote is about education, for example: higher education, or education finance, or schools, or education of underprivileged, or vocational education, or education for handicapped, or excellence, or R&D.
Energy	The quote is about energy, for example: nuclear energy and safety, or electricity, or natural gas & oil, or coal, or alternative and renewable energy, or conservation, or R&D.
Environment	The quote is about the environment, for example: drinking water, or waste disposal, or hazardous waste, or air pollution, or recycling, or species and forest protection, or conservation, or R&D.
Foreign Trade	The quote is about foreign trade, for example: trade agreements, or exports, or private investments, or competitiveness, or tariff and imports, or exchange rates.
Government Operations	The quote is about government operations, for example: intergovernmental relations, or agencies, or bureaucracy, or postal service, or civil employees, or appointments, or national currency, or government procurement, or government property management, or tax administration, or public scandals, or government branch relations, or political campaigns, or census, or capital city, or national holidays.
Health	The quote is about health, for example: health care reform, or health insurance, or drug industry, or medical facilities, or disease prevention, or infants and children, or mental health, or drug/alcohol/tobacco abuse, or R&D.
Housing	The quote is about housing, for example: community development, or urban development, or rural housing, low-income assistance for housing, housing for veterans/elderly/homeless, or R&D.
Immigration	The quote is about migration, for example: immigration, or refugees, or citizenship.
International Affairs	The quote is about international affairs, for example: foreign aid, or international resources exploitation, or developing countries, or international finance, or western Europe, or specific countries, or human rights, or international organisations, or international terrorism, or diplomats.
Labor	The quote is about labour, for example: worker safety, or employment training, or employee benefits, or labor unions, or fair labor standards, or youth employment, or migrant and seasonal workers.
Law and Crime	The quote is about law and crime, for example: law enforcement agencies, or white collar crime, or illegal drugs, or court administration, or prisons, or juvenile crime, or child abuse, or family issues, or criminal and civil code, or police.
Macroeconomics	The quote is about macroeconomics, for example: interest rates, or unemployment, or monetary policy, or national budget, or taxes, or industrial policy.
Public Lands	The quote is about public lands, for example: national parks, or indigenous affairs, or public lands, or water resources, or dependencies and territories.
Social Welfare	The quote is about social welfare, for example: low-income assistance, or elderly assistance, or disabled assistance, or volunteer associations, or child care, or social welfare.

Technology	The quote is about technology, for example: government space programs, or commercial use of space, or science transfer, or telecommunications, or regulation of media, or weather science, or computers, or internet, or R&D.
Transportation	The quote is about transportation, for example: mass transportation, or highways, or air travel, or railroads, or maritime, or infrastructure, or R&D.

Table 17 - CoronaNet hypotheses

label	hypotheses_short	hypotheses_long
Anti-Disinformation Measures	The quote is about measures against disinformation.	The quote is about measures against disinformation: Efforts by the government to limit the spread of false, inaccurate or harmful information.
COVID-19 Vaccines	The quote is about COVID-19 vaccines.	The quote is about COVID-19 vaccines. A policy regarding the research and development, or regulation, or production, or purchase, or distribution of a vaccine..
Closure and Regulation of Schools	The quote is about regulating schools.	The quote is about regulating schools and educational establishments. For example closing an educational institution, or allowing educational institutions to open with or without certain conditions..
Curfew	The quote is about a curfew.	The quote is about a curfew: Domestic freedom of movement is limited during certain times of the day.
Declaration of Emergency	The quote is about declaration of emergency.	The quote is about declaration of a state of national emergency.
External Border Restrictions	The quote is about external border restrictions.	The quote is about external border restrictions: The ability to enter or exit country borders is reduced..
Health Monitoring	The quote is about health monitoring.	The quote is about health monitoring of individuals who are likely to be infected..
Health Resources	The quote is about health resources, materials, infrastructure, personnel, mask purchases.	The quote is about health resources: For example medical equipment, number of hospitals, health infrastructure, personnel (e.g. doctors, nurses), mask purchases.
Health Testing	The quote is about health testing.	The quote is about health testing of large populations regardless of their likelihood of being infected..
Hygiene	The quote is about hygiene.	The quote is about hygiene: Promotion of hygiene in public spaces, for example disinfection in subways or burials..
Internal Border Restrictions	The quote is about internal border restrictions.	The quote is about internal border restrictions: The ability to move freely within the borders of a country is reduced..
Lockdown	The quote is about a lockdown.	The quote is about a lockdown: People are obliged shelter in place and are only allowed to leave their shelter for specific reasons.
New Task Force, Bureau or	The quote is about a new administrative body.	The quote is about a new administrative body, for example a new task force, bureau or administrative configuration..

Administrative Configuration		
Public Awareness Measures	The quote is about public awareness measures.	The quote is about public awareness measures or efforts to disseminate or gather reliable information, for example information on health prevention..
Quarantine	The quote is about quarantine.	The quote is about quarantine. People are obliged to isolate themselves if they are infected..
Restriction and Regulation of Businesses	The quote is about restricting or regulating businesses.	The quote is about restricting or regulating businesses, private commercial activities: For example closing down commercial establishments, or allowing commercial establishments to open with or without certain conditions..
Restriction and Regulation of Government Services	The quote is about restricting or regulating government services or public facilities.	The quote is about restricting or regulating government services or public facilities: For example closing down government services, or allowing government services to operate with or without certain conditions..
Restrictions of Mass Gatherings	The quote is about restrictions of mass gatherings.	The quote is about restrictions of mass gatherings: The number of people allowed to congregate in a place is limited.
Social Distancing	The quote is about social distancing, reducing contact, mask wearing.	The quote is about social distancing, reducing contact between individuals in public spaces, mask wearing..
Other Policy Not Listed Above	The quote is about something other than regulation of businesses, government, gatherings, distancing, quarantine, lockdown, curfew, emergency, vaccine, disinformation, schools, borders or travel, testing, resources. It is not about any of these topics..	The quote is about something other than regulation of businesses, government, gatherings, distancing, quarantine, lockdown, curfew, emergency, vaccines, disinformation, schools, borders or travel, testing, health resources. It is not about any of these topics..

Appendix C: Analysis Pipeline

To ensure comparability across algorithms and datasets as well as reproducibility, each dataset was analysed with the following overall steps.[12]

1. Train-test-split: given a *dataset,* create a training set *data_train* and a held-out test set *data_test*. The train-test-split proportions depend on the dataset, see appendix A.
2. Random sampling: From *data_train*, take a fully random sample *data_train_samp* of size *N*.
3. Hyperparameter tuning with cross-validation: Determine the best hyperparameters *hyperparam_best* for the algorithm on *data_train_samp*. We do not assume access to a development/validation set and therefore use two-fold cross-validation to find *hyperparam_best*. We use the Python library Optuna[13] for smart sampling of the best hyperparameters. We search over up to 60 hyperparameter configurations for the classical algorithm and up to 23 for Transformers, given their high computational training costs. For each hyperparameter configuration, step 2 and 3 are repeated twice for two random seeds to account for randomness in sampling *data_train_samp*.
4. Training: Use *data_train_samp* and *hyperparam_best* to train the algorithm *algo*.
5. Testing: Test *algo* on *data_test* using metrics F1-micro and F1-macro.
6. Account for randomness: Repeat step 4 and 5 three times with three different random seeds for sampling *data_train_samp*. Calculate the mean F1-micro and F1-macro as well as standard deviation to account for the impact of randomness on performance.
7. Repeat for different sample sizes: Repeat steps 2 to 7 for each *N* in *[0, 100, 500, 1000, 2500, 5000, 10 000]* to test the performance of *algo* depending on the number of training examples.
8. Repeat for different algorithms: Repeat steps 2 to 8 for each *algo* in *[SVM, Logistic Regression, BERT-base, BERT-NLI].* The steps are repeated twice for SVM and Logistic Regression, once with TFIDF vectorization and once with averaged word embeddings (see details in appendix F on pre-processing).
9. Repeat for different datasets: Repeat steps 1 to 9 for each *dataset* in *[sentiment-economy, CoronaNet, Manifesto-8-class, CAP-SotU, CAP-US-Court, Manifesto-Military, Manifesto-Protectionism, Manifesto-Morality]*

[12] The full script written in Python is available on our GitHub repository: https://github.com/MoritzLaurer/less-annotating-with-bert-nli

[13] https://optuna.readthedocs.io/en/stable/

Appendix D: Metrics per algorithm per sample size

The following figures and tables display the exact metrics underlying the text and figures in the paper. We start with a comparison of different metrics and what they tell us about how different algorithms can handle imbalanced data. Based on this comparison, we show the advantages and disadvantages of different metrics and argue why F1-macro is the best metric for many social science use-cases. The tables in the following sub-section D3 then display the average metrics across all datasets (or tasks) for each sample size interval and algorithm. The tables in the final subsection D4 display the metrics for each dataset individually.

D1: Comparison across metrics: how well can different algorithms handle data imbalance?

Assuming that all classes in a task have similar substantive value, a good algorithm should perform similarly across on all classes with little deviation. We assess this characteristic by first calculating metrics for each class individually and then calculating the standard deviation across classes. Lower standard deviation indicates more similar performance across classes. Figure 3 shows the cross-class standard deviation for all datasets averaged across the data intervals 100 to 2500. Cross-class standard deviation is lowest for BERT-NLI and highest for classical algorithms with TFIDF. This data supports the claim in the main text that more transfer learning increases the ability of classifiers to handle imbalanced data and predict minority classes.

Moreover, we argue that higher standard deviation mostly stems from overpredicting a few majority classes, while underpredicting the remaining classes. Figure 4 shows the average performance of each algorithm on all classes (left column); performance on the top 25% of classes with the most data (middle column, this includes e.g. the 5 ‘majority classes’ for a task with 20 classes in total); and performance on the bottom 75% of classes in terms of number

of data points (right column). This figure shows that BERT-NLI performs comparatively worse on the top 25% classes and comparatively well on the bottom 75% of classes. This analysis shows empirically, that BERT-NLI favours (many) minority classes over (few) majority classes. BERT-base and especially classical algorithms base their aggregate performance more strongly on favouring (few) majority classes and disfavouring (many) minority classes. The opposite tendency can be observed for precision. BERT-NLI is more precise for majority classes (it predicts them less, but more precisely i.e. with less false positives), but it is less precise for the remaining classes (it predicts them more, but less precisely i.e. with more false positives). Which variant is better, depends on the substantive use-case. On average, we assume that equal performance across classes independently of their size is best for many social science use-cases.

D2: Which metric is most adequate for social science use-cases?

Based on this empirical comparison of metrics, we conclude that F1-macro is the best metric for 'average social science use-cases'. It weighs all classes equally and provides the harmonized mean of precision and recall. Its main disadvantage is that it is less straight forward to interpret than accuracy. Balanced accuracy provides a more interpretable alternative, with the disadvantages of neglecting precision (it is equivalent to recall-macro). It favours classifiers overpredicting minority classes and underpredicting majority classes.

Figure 5 shows the average performance of all algorithms on several additional metrics: recall-macro, recall-micro, precision-macro, precision-micro, Cohen's Kappa and Matthews correlation coefficient. The figure shows that the overall tendencies for all metrics are the

same: the two BERT variants clearly outperform the classical algorithms. As more data is added, BERT-NLI and BERT-base become aligned.

We also believe that standard accuracy or any other *-micro metric should only be used as a primary metric in very few use-cases where data is fully balanced. We did not find a single balanced social science dataset. For example: CAP SotU has 3114 texts on international affairs and 235 on Immigration. We assume that in most substantive use-cases, international affairs is not 13.2 times more important than immigration; CoronaNet has 6468 texts on Health Resources and 419 texts on Anti-Disinformation Measures. Health Resources is probably not 15.4 times more important in most use-cases; Manifesto-military has 590 texts that are negative towards the military and more than 10655 texts that are about something else. Texts about something other than the military are not 18 times more important for this task. Metrics like accuracy, recall-micro or precision-micro literally make the assumption that these majority classes are 13.2/15.4/18 times more important – which seems to be wrong in many substantive use-cases.

Figure 3 - Cross-class standard deviation averaged across all datasets

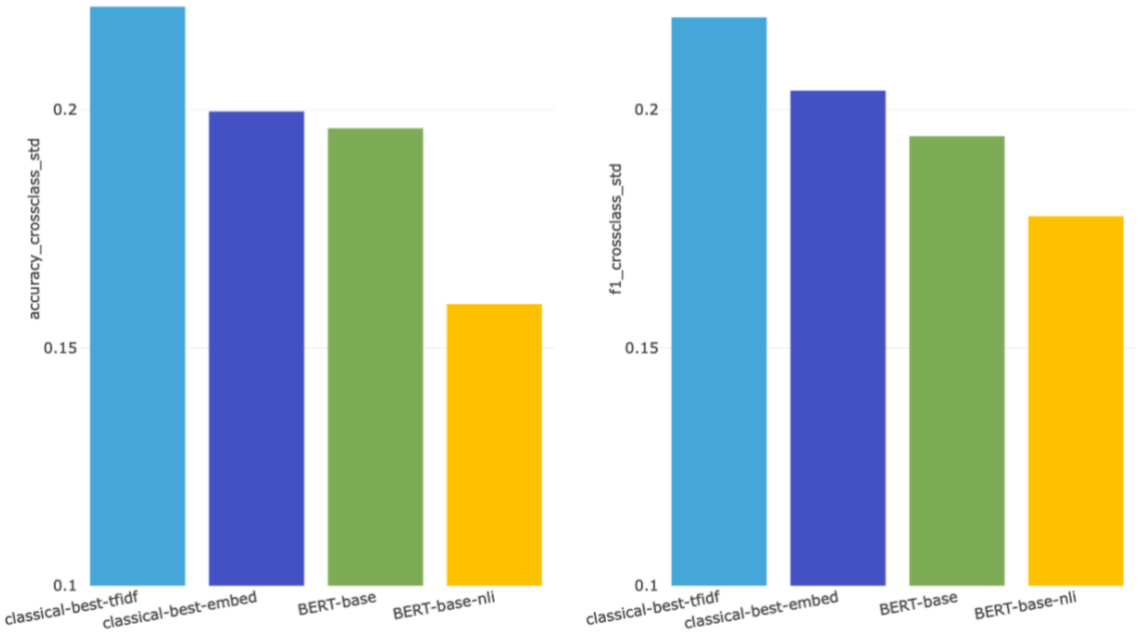


Figure 4 - Comparison of averaged metrics for top 25% classes and bottom 75% classes

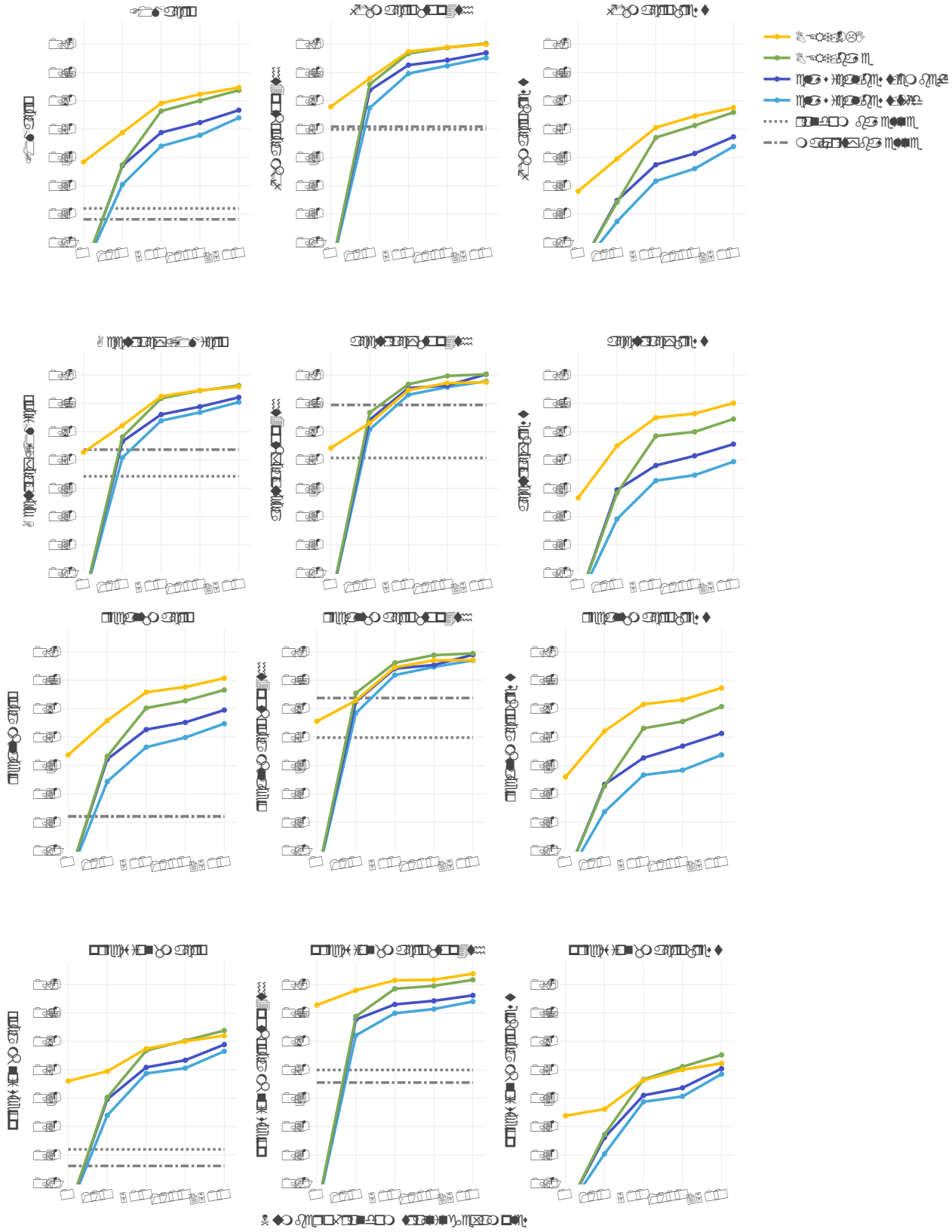
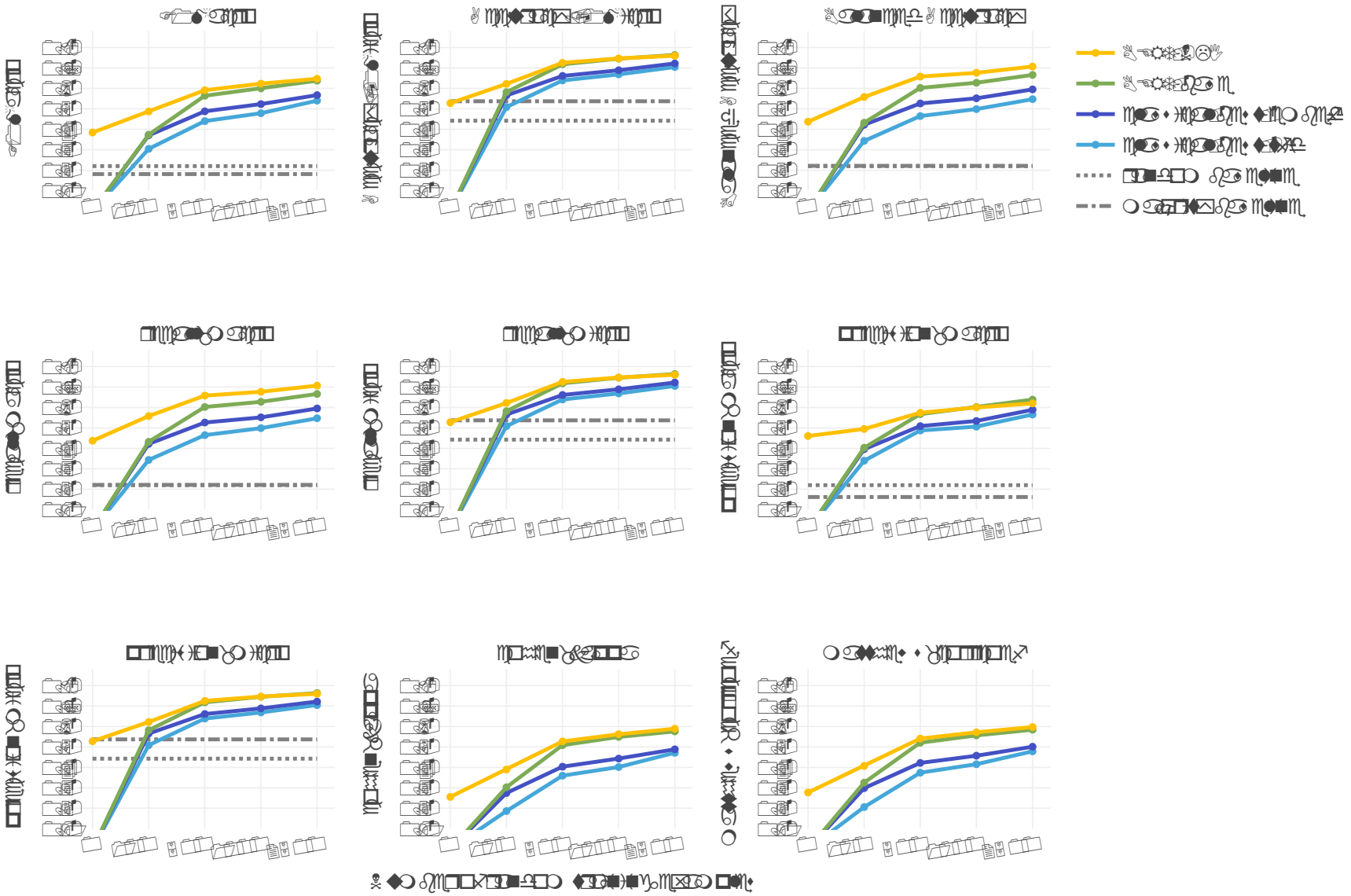


Figure 5 - Aggregate scores averaged across all datasets for many different metrics



D3: Aggregated Metrics Across Datasets

The tables below display the average metrics across all datasets (or tasks) for each sample size interval and algorithm. For example, the column "2500 (8 datasets)" displays the average metrics for all eight datasets (tasks) with training data sample size 2500. Note that for the intervals 5000 and 10000 only four and three datasets had sufficient data. The metrics after 2500 are therefore not directly comparable with previous intervals. Also note that "BERT" always refers to DeBERTaV3 (see appendix B). We have underlined the values used in the main text to make it easier for readers to link this data to the main text.

Table 18 - Average F1-macro across all datasets

Sample size / Algorithm	0 (8 datasets)	100 (8 datasets)	500 (8 datasets)	1000 (8 datasets)	2500 (8 datasets)	5000 (4 datasets)	10000 (3 datasets)
SVM_tfidf	0	0,285	0,44	0,478	0,54	0,469	0,486
logistic_tfidf	0	0,304	0,434	0,465	0,516	0,455	0,478
SVM_embeddings	0	0,355	0,469	0,516	0,567	0,538	0,56
logistic_embeddings	0	0,37	0,488	0,523	0,563	0,532	0,554
classical-best-tfidf	0	0,304	0,44	0,478	0,54	0,469	0,486
classical-best-embed	0	0,37	0,488	0,523	0,567	0,538	0,56
BERT-base	0	0,374	0,564	0,6	0,637	0,621	0,65
BERT-base-nli	0,384	0,487	0,591	0,623	0,647	0,597	0,626

Table 19 - Average F1-micro/accuracy across all datasets

Sample size / Algorithm	0 (8 datasets)	100 (8 datasets)	500 (8 datasets)	1000 (8 datasets)	2500 (8 datasets)	5000 (4 datasets)	10000 (3 datasets)
SVM_tfidf	0	0,508	0,639	0,669	0,704	0,579	0,584
logistic_tfidf	0	0,507	0,623	0,649	0,684	0,562	0,576
SVM_embeddings	0	0,557	0,66	0,678	0,71	0,621	0,617
logistic_embeddings	0	0,565	0,661	0,689	0,722	0,621	0,622
classical-best-tfidf	0	0,508	0,639	0,669	0,704	0,579	0,584
classical-best-embed	0	0,565	0,661	0,689	0,722	0,621	0,622
BERT-base	0	0,582	0,718	0,745	0,763	0,691	0,699
BERT-base-nli	0,528	0,622	0,725	0,746	0,759	0,672	0,677

Table 20 - Average balanced accuracy across all datasets

Sample size / Algorithm	0 (8 datasets)	100 (8 datasets)	500 (8 datasets)	1000 (8 datasets)	2500 (8 datasets)	5000 (4 datasets)	10000 (3 datasets)
SVM_tfidf	0	0,321	0,458	0,498	0,547	0,465	0,464
logistic_tfidf	0	0,343	0,464	0,491	0,535	0,449	0,464
SVM_embeddings	0	0,402	0,509	0,552	0,595	0,524	0,555
logistic_embeddings	0	0,422	0,526	0,551	0,582	0,514	0,536
classical-best-tfidf	0	0,343	0,464	0,498	0,547	0,465	0,464
classical-best-embed	0	0,422	0,526	0,552	0,595	0,524	0,555
BERT-base	0	0,432	0,602	0,627	0,666	0,625	0,651
BERT-base-nli	0,437	0,558	0,658	0,676	0,707	0,642	0,66

Table 21 - Average F1-macro difference between different algorithms, all datasets

Sample size	0 (8 datasets)	100 (8 datasets)	500 (8 datasets)	1000 (8 datasets)	2500 (8 datasets)	5000 (4 datasets)	10000 (3 datasets)	mean (100 to 2500)	mean all
classical-best-embed vs. classical-best-tfidf	0	0,066	0,048	0,045	0,027	0,069	0,074	0,046	0,055
BERT-base vs. classical-best-tfidf	0	0,07	0,124	0,122	0,097	0,152	0,164	0,103	0,122
BERT-base vs. classical-best-embed	0	0,004	0,076	0,077	0,07	0,083	0,09	0,057	0,067
BERT-base-nli vs. classical-best-tfidf	0,384	0,183	0,151	0,145	0,107	0,128	0,14	0,146	0,142
BERT-base-nli vs. classical-best-embed	0,384	0,117	0,103	0,1	0,08	0,059	0,066	0,1	0,087
BERT-base-nli vs. BERT-base	0,384	0,113	0,027	0,023	0,01	-0,024	-0,024	0,043	0,021

Table 22 - Average F1-micro difference between different algorithms, all datasets

Sample size	0 (8 datasets)	100 (8 datasets)	500 (8 datasets)	1000 (8 datasets)	2500 (8 datasets)	5000 (4 datasets)	10000 (3 datasets)	mean (100 to 2500)	mean all
classical-best-embed vs. classical-best-tfidf	0	0,057	0,022	0,02	0,018	0,042	0,038	0,029	0,033
BERT-base vs. classical-best-tfidf	0	0,074	0,079	0,076	0,059	0,112	0,115	0,072	0,086
BERT-base vs. classical-best-embed	0	0,017	0,057	0,056	0,041	0,07	0,077	0,043	0,053
BERT-base-nli vs. classical-best-tfidf	0,528	0,114	0,086	0,077	0,055	0,093	0,093	0,083	0,086
BERT-base-nli vs. classical-best-embed	0,528	0,057	0,064	0,057	0,037	0,051	0,055	0,054	0,054
BERT-base-nli vs. BERT-base	0,528	0,04	0,007	0,001	-0,004	-0,019	-0,022	0,011	0,001

Table 23 - Average balanced accuracy difference between different algorithms, all datasets

Sample size	0 (8 datasets)	100 (8 datasets)	500 (8 datasets)	1000 (8 datasets)	2500 (8 datasets)	5000 (4 datasets)	10000 (3 datasets)	mean (100 to 2500)	mean all
classical-best-embed vs. classical-best-tfidf	0	0,079	0,062	0,054	0,048	0,059	0,091	0,061	0,065
BERT-base vs. classical-best-tfidf	0	0,089	0,138	0,129	0,119	0,16	0,187	0,119	0,137
BERT-base vs. classical-best-embed	0	0,01	0,076	0,075	0,071	0,101	0,096	0,058	0,072
BERT-base-nli vs. classical-best-tfidf	0,437	0,215	0,194	0,178	0,16	0,177	0,196	0,187	0,187
BERT-base-nli vs. classical-best-embed	0,437	0,136	0,132	0,124	0,112	0,118	0,105	0,126	0,121
BERT-base-nli vs. BERT-base	0,437	0,126	0,056	0,049	0,041	0,017	0,009	0,068	0,05

Table 24 - Average F1-macro across four datasets with 5000 data points or more

Sample size / Algorithm	0 (8 datasets)	100 (8 datasets)	500 (8 datasets)	1000 (8 datasets)	2500 (8 datasets)	5000 (4 datasets)	10000 (3 datasets)
SVM_tfidf	0	0,126	0,329	0,387	0,441	0,469	0,492
logistic_tfidf	0	0,168	0,342	0,388	0,43	0,455	0,484
SVM_embeddings	0	0,228	0,389	0,436	0,505	0,538	0,558
logistic_embeddings	0	0,25	0,398	0,447	0,503	0,532	0,55
classical-best-tfidf	0	0,168	0,342	0,388	0,441	0,469	0,492
classical-best-embed	0	0,25	0,398	0,447	0,505	0,538	0,558
BERT-base	0	0,228	0,478	0,523	0,582	0,621	0,636
BERT-base-nli	0,281	0,417	0,523	0,549	0,567	0,597	0,612

Table 25 - Average F1-micro across four datasets with 5000 data points or more

Sample size / Algorithm	0 (8 datasets)	100 (8 datasets)	500 (8 datasets)	1000 (8 datasets)	2500 (8 datasets)	5000 (4 datasets)	10000 (3 datasets)
SVM_tfidf	0	0,302	0,474	0,522	0,557	0,579	0,604
logistic_tfidf	0	0,322	0,483	0,51	0,543	0,562	0,598
SVM_embeddings	0	0,394	0,525	0,55	0,594	0,621	0,636
logistic_embeddings	0	0,41	0,519	0,552	0,609	0,621	0,637
classical-best-tfidf	0	0,322	0,483	0,522	0,557	0,579	0,604
classical-best-embed	0	0,41	0,525	0,552	0,609	0,621	0,637
BERT-base	0	0,398	0,589	0,625	0,662	0,691	0,703
BERT-base-nli	0,305	0,474	0,602	0,628	0,642	0,672	0,684

Table 26 - Average balanced accuracy across four datasets with 5000 data points or more

Sample size / Algorithm	0 (8 datasets)	100 (8 datasets)	500 (8 datasets)	1000 (8 datasets)	2500 (8 datasets)	5000 (4 datasets)	10000 (3 datasets)
SVM_tfidf	0	0,138	0,313	0,371	0,435	0,465	0,477
logistic_tfidf	0	0,17	0,33	0,375	0,418	0,449	0,468
SVM_embeddings	0	0,233	0,378	0,426	0,491	0,524	0,545
logistic_embeddings	0	0,256	0,392	0,439	0,477	0,514	0,529
classical-best-tfidf	0	0,17	0,33	0,375	0,435	0,465	0,477
classical-best-embed	0	0,256	0,392	0,439	0,491	0,524	0,545
BERT-base	0	0,246	0,48	0,522	0,583	0,625	0,633
BERT-base-nli	0,359	0,47	0,554	0,584	0,62	0,642	0,653

D4: Disaggregated Metrics Per Dataset

The tables below display the metrics for each dataset individually. Note that a single metric in one cell for one algorithm on one dataset for one sample size is itself the average of three different random runs with different random seeds for sampling and parameter initialisation. The combination of many datasets/tasks, sample size intervals, algorithms and random seeds is designed to create a robust estimate of the general performance of different approaches. Note that for example "f1_macro_mean" refers to the mean of three different F1-macro results for three different random seeds. Since we have calculated over a dozen different metrics for each dataset, we cannot display all of them here. Much more extensive data, with many more metrics and standard deviations for each metric can be found in Excel files in our GitHub repository in the appendix folder.[14]

[14] https://github.com/MoritzLaurer/less-annotating-with-bert-nli/tree/master/appendix

Table 27 - Manifesto-8 detailed metrics

	n_sample	logistic_tfidf	SVM_tfidf	logistic_embeddings	SVM_embeddings	deberta-v3-base	deberta-v3-nli
f1_macro_mean	0	0	0	0	0	0	0,059[15]
	100	0,171	0,116	0,276	0,286	0,216	0,327
	500	0,286	0,305	0,374	0,387	0,383	0,451
	1000	0,31	0,329	0,411	0,408	0,438	0,474
	2500	0,355	0,368	0,44	0,434	0,471	0,49
	5000	0,384	0,387	0,453	0,45	0,497	0,508
	10000	0,404	0,415	0,467	0,464	0,525	0,521
accuracy /f1_micro_mean	0	0	0	0	0	0	0,036
	100	0,257	0,268	0,4	0,402	0,33	0,388
	500	0,416	0,43	0,494	0,524	0,512	0,548
	1000	0,42	0,463	0,539	0,526	0,567	0,581
	2500	0,486	0,51	0,58	0,55	0,589	0,597
	5000	0,52	0,535	0,589	0,569	0,619	0,614
	10000	0,541	0,555	0,598	0,58	0,644	0,629

Table 28 - Manifesto-military detailed metrics

	n_sample	logistic_tfidf	SVM_tfidf	logistic_embeddings	SVM_embeddings	deberta-v3-base	deberta-v3-nli
f1_macro_mean	0	0	0	0	0	0	0,592
	100	0,341	0,358	0,497	0,491	0,547	0,649
	500	0,534	0,565	0,58	0,603	0,656	0,698
	1000	0,568	0,574	0,615	0,632	0,67	0,725
	2500	0,622	0,636	0,645	0,645	0,685	0,738
	3970 (all)	0,626	0,668	0,652	0,672	0,704	0,755
accuracy /f1_micro_mean	0	0	0	0	0	0	0,924
	100	0,61	0,676	0,798	0,8	0,858	0,905
	500	0,842	0,874	0,872	0,889	0,927	0,929
	1000	0,877	0,873	0,898	0,899	0,929	0,935
	2500	0,906	0,907	0,913	0,906	0,929	0,938
	3970 (all)	0,91	0,934	0,914	0,913	0,929	0,939

[15] Note that this low performance of BERT-NLI on Manifesto-8 with zero training data is not a coding error. Without task-specific training data, the NLI model systematically over-estimated the "Not other category applies" class the corresponding class-hypothesis "The quote is about something other than the topics economy, international relations, society, freedom and democracy, political system, social groups, welfare. It is about non of these topics." This issue could have been remedied by formulating a different hypothesis.

Table 29 - Manifesto-protectionism detailed metrics

	n_sample	logistic_tfidf	SVM_tfidf	logistic_embeddings	SVM_embeddings	deberta-v3-base	deberta-v3-nli
f1_macro_mean	0	0	0	0	0	0	0,379
	100	0,492	0,509	0,518	0,504	0,529	0,465
	500	0,542	0,566	0,611	0,586	0,667	0,646
	1000	0,516	0,594	0,638	0,616	0,681	0,715
	2116 (all)	0,626	0,672	0,628	0,646	0,675	0,737
accuracy /f1_micro_mean	0	0	0	0	0	0	0,708
	100	0,804	0,859	0,767	0,766	0,772	0,733
	500	0,826	0,872	0,862	0,844	0,875	0,872
	1000	0,809	0,885	0,887	0,871	0,908	0,912
	2116 (all)	0,891	0,919	0,878	0,883	0,886	0,918

Table 30 - Manifesto-morality detailed metrics

	n_sample	logistic_tfidf	SVM_tfidf	logistic_embeddings	SVM_embeddings	deberta-v3-base	deberta-v3-nli
f1_macro_mean	0	0	0	0	0	0	0,337
	100	0,405	0,393	0,391	0,387	0,429	0,43
	500	0,456	0,49	0,507	0,5	0,591	0,574
	1000	0,483	0,501	0,545	0,553	0,642	0,646
	2500	0,502	0,607	0,611	0,615	0,708	0,712
	3188 (all)	0,564	0,617	0,61	0,632	0,713	0,708
accuracy /f1_micro_mean	0	0	0	0	0	0	0,679
	100	0,759	0,74	0,718	0,721	0,784	0,742
	500	0,775	0,804	0,792	0,786	0,873	0,854
	1000	0,803	0,82	0,838	0,829	0,89	0,887
	2500	0,803	0,883	0,867	0,858	0,922	0,911
	3188 (all)	0,853	0,886	0,868	0,867	0,914	0,912

Table 31 - Sentiment economy news detailed metrics

	n_sample	logistic_tfidf	SVM_tfidf	logistic_embeddings	SVM_embeddings	deberta-v3-base	deberta-v3-nli
f1_macro_mean	0	0	0	0	0	0	0,638
	100	0,522	0,517	0,56	0,546	0,571	0,684
	500	0,57	0,581	0,61	0,507	0,685	0,717

	1000	0,601	0,609	0,597	0,582	0,712	0,704
	2500	0,654	0,64	0,611	0,61	0,701	0,723
	3000 (all)	0,674	0,655	0,621	0,591	0,702	0,727
accuracy /f1_micro_mean	0	0	0	0	0	0	0,688
	100	0,592	0,584	0,599	0,595	0,645	0,698
	500	0,61	0,662	0,684	0,661	0,71	0,736
	1000	0,662	0,682	0,68	0,625	0,735	0,723
	2500	0,703	0,699	0,681	0,66	0,721	0,741
	3000 (all)	0,712	0,708	0,683	0,668	0,721	0,745

Table 32 - CAP state of the union detailed metrics

	n_sample	logistic_tfidf	SVM_tfidf	logistic_embeddings	SVM_embeddings	deberta-v3-base	deberta-v3-nli
f1_macro_mean	0	0	0	0	0	0	0,36
	100	0,181	0,113	0,218	0,179	0,197	0,448
	500	0,319	0,291	0,384	0,352	0,426	0,536
	1000	0,361	0,369	0,43	0,423	0,491	0,563
	2500	0,38	0,408	0,501	0,516	0,624	0,567
	5000	0,38	0,415	0,547	0,547	0,661	0,626
	10000	0,453	0,443	0,572	0,576	0,687	0,662
accuracy /f1_micro_mean	0	0	0	0	0	0	0,425
	100	0,342	0,297	0,386	0,36	0,415	0,52
	500	0,454	0,434	0,501	0,508	0,545	0,591
	1000	0,469	0,49	0,541	0,537	0,586	0,604
	2500	0,475	0,513	0,592	0,584	0,663	0,618
	5000	0,476	0,513	0,6	0,59	0,691	0,673
	10000	0,566	0,563	0,626	0,612	0,706	0,692

Table 33 - CAP US court cases detailed metrics

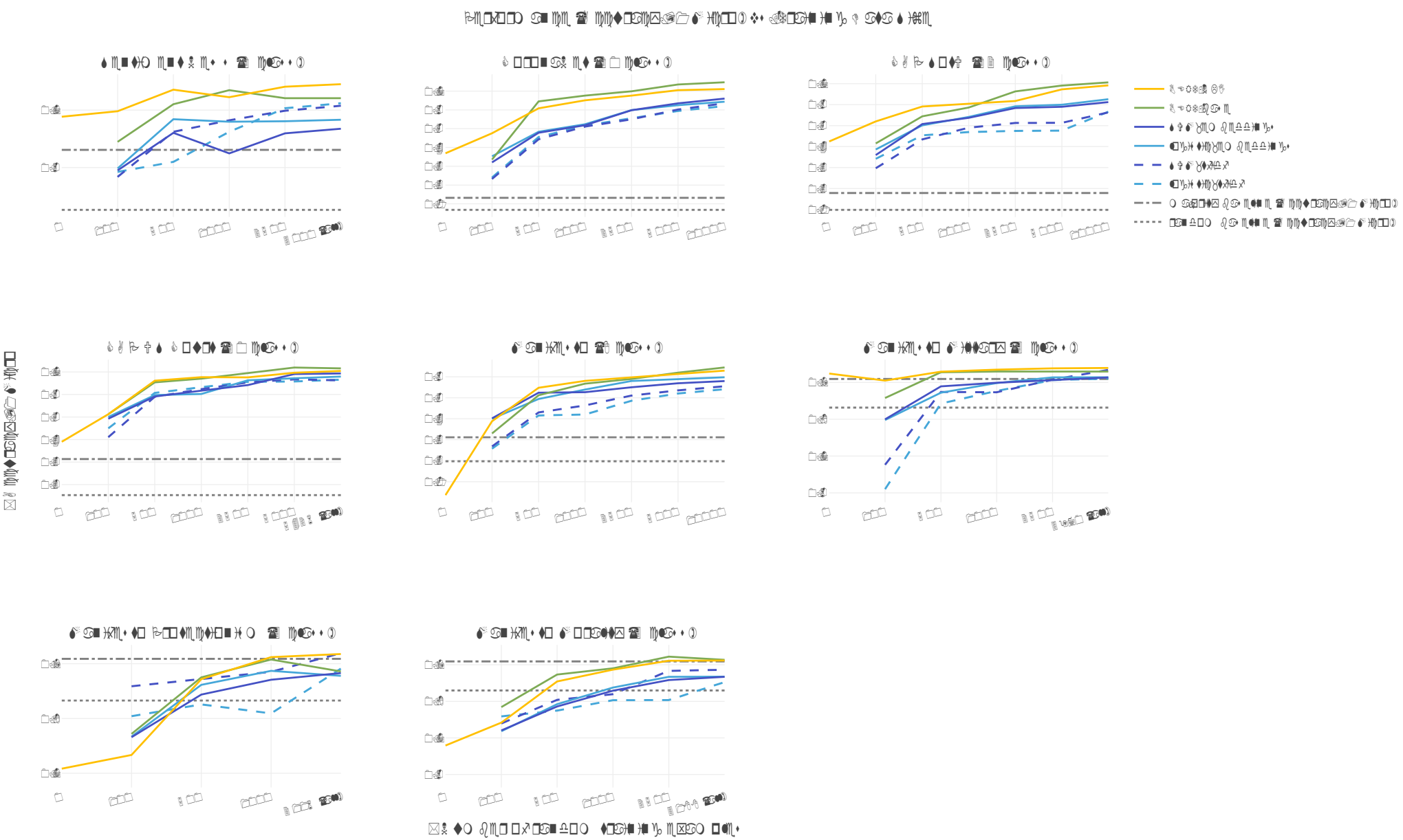
	n_sample	logistic_tfidf	SVM_tfidf	logistic_embeddings	SVM_embeddings	deberta-v3-base	deberta-v3-nli
f1_macro_mean	0	0	0	0	0	0	0,32
	100	0,141	0,112	0,229	0,197	0,23	0,422
	500	0,353	0,323	0,402	0,377	0,483	0,522
	1000	0,412	0,39	0,453	0,43	0,515	0,533
	2500	0,471	0,473	0,508	0,505	0,551	0,557
	5000	0,506	0,511	0,527	0,545	0,603	0,57
	5426 (all)	0,498	0,508	0,538	0,555	0,593	0,57

accuracy /f1_micro_mean	0	0	0	0	0	0	0,39
	100	0,45	0,411	0,501	0,492	0,513	0,513
	500	0,607	0,59	0,597	0,591	0,654	0,662
	1000	0,633	0,624	0,602	0,617	0,671	0,677
	2500	0,656	0,654	0,663	0,642	0,697	0,676
	5000	0,658	0,667	0,672	0,691	0,72	0,697
	5426 (all)	0,666	0,663	0,68	0,693	0,716	0,705

Table 34 - CoronaNet detailed metrics

	n_sample	**logistic_tfidf**	**SVM_tfidf**	**logistic_embeddings**	**SVM_embeddings**	**deberta-v3-base**	**deberta-v3-nli**
f1_macro_mean	0	0	0	0	0	0	0,386
	100	0,18	0,165	0,275	0,248	0,27	0,472
	500	0,411	0,397	0,433	0,439	0,619	0,585
	1000	0,467	0,46	0,496	0,481	0,649	0,626
	2500	0,516	0,514	0,562	0,564	0,683	0,652
	5000	0,551	0,562	0,6	0,609	0,722	0,685
	10000	0,578	0,599	0,622	0,638	0,738	0,695
accuracy /f1_micro_mean	0	0	0	0	0	0	0,369
	100	0,24	0,232	0,353	0,32	0,336	0,476
	500	0,455	0,444	0,484	0,478	0,645	0,608
	1000	0,517	0,511	0,524	0,519	0,676	0,651
	2500	0,554	0,55	0,599	0,598	0,699	0,676
	5000	0,594	0,602	0,625	0,635	0,735	0,705
	10000	0,62	0,635	0,644	0,66	0,747	0,711

Figure 6 – Accuracy / F1-micro performance per dataset per sample size per algorithm



Appendix E: Pre-processing and Hyperparameters

The tables below display the best hyperparameters for each dataset and sample size determined by a hyperparameter search with the Python library Optuna[16] for all algorithms. Optuna starts with a random hyperparameter search and then smartly samples well performing hyperparameters to avoid the high computational costs of grid search. For classical algorithms 60 different hyperparameter configurations were tested, for BERT-nli up to 23 configurations and for BERT-base up to 15 configurations were tested to save computational resources. We tested more configurations for BERT-NLI, as it has two additional hyperparameters (warmup ratio and the hypothesis formulation) and it is faster to train than BERT-base because it needs less epochs for good performance. Moreover, for BERT models no hyperparameter search was conducted for sample size 10 000 (and 5000 for some datasets with very long texts). Hyperparameters for sample size 5000 were used instead, because optimal hyperparameters seemed to stay relatively stable across sample sizes and to avoid high computation costs for minimal performance benefits. Note that we call the algorithms "BERT" for simplicity, the actual pre-trained algorithm used was DeBERTaV3-base (He, Gao, and Chen 2021).[17]

We also discuss pre-processing decisions below, as several pre-processing options were treated as hyperparameters during hyperparameter search to test optimal pre-processing methods.

[16] https://optuna.readthedocs.io/en/stable/

[17] DeBERTaV3-base can be downloaded at https://huggingface.co/microsoft/deberta-v3-base. Our DeBERTa-nli can be downloaded at https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-docnli-ling-2c

E1: Include context sentences or not?

For quasi-sentence level datasets (Manifesto datasets and State of the Union Speeches), we tested concatenating the target quasi-sentence with its preceding and following quasi-sentences. The underlying assumption is that these context sentences provided additional relevant information to the classifier (Bilbao-Jayo and Almeida 2018). We find that including context systematically increases performance.[18]

This has probably two main reasons: First, the surrounding sentences contain relevant information for the classifier to contextualise the target sentence. Humans also annotated each sentence after reading its context instead of isolated strings and would most likely also perform worse if no context were provided. Second, we find that especially for the CAP-SotU datasets it is also simply statistically likely that the surrounding sentences have the same class as the target sentence. Including the surrounding sentences can therefore also be an effective means for data augmentation, depending on the dataset and task. The method of including surrounding sentences needs to be used carefully though, as researchers need to make sure that context sentences are not used in both the training and test dataset. We therefore implemented the train-test-split for these datasets on the document level instead of the sentence level to avoid data leakage from the training set into the test set.

Table 35 – Likelihood of surrounding sentences having the same label as the target sentence

	CAP-SotU	**Manifesto-8**
Preceding quasi-sentence has same label as target quasi-sentence	75.4%	57.4%
Following quasi-sentence has same label as target quasi-sentence	75.5%	57.4%

[18] If the word "context" is part of the string in the column "context" in the tables below, the hyperparameter search determined that including the context sentences is beneficial for performance.

E2: Pre-processing for BERT and BERT-NLI

One advantage of BERT-base algorithms is that text pre-processing is simple. Each algorithm comes with its own tokenizer, which converts words to sub-word tokens, which can then be processed by the algorithm. Operations like stop-word removal or lemmatization are not necessary, as the algorithms' vocabulary covers all possible (sub-)words and it automatically down-weights unimportant words. Out-of-vocabulary issues are not possible, as very rare words or typos would be automatically converted into sub-strings/individual characters by the tokenizer. For BERT-NLI, however, special pre-processing is necessary to create the necessary hypothesis-context input format. Appendix B provides details and examples for the pre-processing steps for BERT-NLI. For BERT-NLI, the formulation of the hypothesis becomes a new hyperparameter, which we tested during hyperparameter search (see the hyperparameter table below and appendix B).

E3: Choosing hyperparameters – advice for BERT models

Choosing the right hyperparameters can be a challenge when starting to work with Transformers like BERT. We therefore provide advice based on extensive experiments. We first conducted some initial tests, after which we discarded some hyperparameters such as learning rate decay, or dropout as we did not notice relevant impacts on performance. We then focussed our extensive hyperparameter search on three main hyperparameters: Learning rate, epochs, and batch size. For BERT-NLI, we noticed that warmup-ration is another useful hyperparameter (see details below). The following advice is based on our experience

and on the hyperparameter importance scores created by the Optuna library (see tables below).

- One **epoch** is one iteration over the entire training dataset. More epochs enable to algorithm to learn more from the dataset, while too many epochs risk overfitting to the specific training set and reducing performance on holdout test sets. With larger datasets, BERT is normally only trained for up to 10 epochs. We noticed, however, that performance continued increasing with very high epochs, especially for smaller sample sizes. We therefore tested epochs in the range of {30, 100} for BERT-base and up to 50 for BERT-NLI as BERT-NLI does not need to learn the task from scratch. Our experiments show that many epochs can still increase performance. Very high numbers of epochs cost, however, more computation and improvements are only marginal. As lower epochs also led to optimal hyperparameters for several datasets, we recommend training for up to 40 epochs for data sizes smaller than 10000. For BERT-NLI, much less epochs also lead to good performance (e.g. around 5). For datasets larger than 10000, less than 10 epochs can be used, but training for up to around 40 epochs may lead to improved performance with BERT-base.
- The **batch size** determines the number of annotated texts the algorithm sees until its internal parameters are updated. With a batch size of 16, the algorithm sees 16 annotated texts before it updates its parameters to 'learn' from these texts (one 'training step'). Overall, the batch size did not have a very important impact of performance. Based on our experience, we recommend the following: If the dataset is very small (around 100), a smaller batch size (8 or 16) can be helpful to ensure enough batches. If the dataset gets larger, the importance of batch size seems to diminish. Advantages of higher batch sizes are increases in computational speed and an increased likelihood that a batch includes

smaller classes for imbalanced datasets. Our general recommendation is therefore to use larger batch sizes (16 or 32) especially as dataset size grows and if GPU memory permits.

- The **learning rate** is the most important hyperparameter. It determines how strongly the algorithm's parameters are updated with each batch. With a high learning rate, the parameters are updated strongly with each batch, risking overfitting to batches. With a low learning rate, the parameters are only updated a little, risking that the algorithm does not 'learn' enough. We tested learning rates mostly in the range of {1e-6, 5e-4}. We find that the optimal learning rate is mostly close to standard learning rates of around 2e-5 which is also recommended in (He, Gao, and Chen 2021). We conclude that a hyperparameter search between 9e-6 to 4e-5 is sufficient. In case of limited computational resources, we also assume that choosing a default learning rate of 2e-5 will lead to good performance in most cases and resources for hyperparameter searches can be saved.
- **Warmup ratio** is another hyperparameter that turned out to be relevant for BERT-NLI, but is less important for standard BERT models. With a warmup ratio of e.g. 0.4, the learning rate first starts at 0 during training and then linearly increases to the full learning rate after the first 40% of training steps. As BERT-NLI already starts with relevant task knowledge, a higher warmup ratio avoids that BERT-NLI forgets relevant task knowledge during the first training steps with a too high learning rate (a phenomenon called 'catastrophic forgetting'). For BERT-base, a standard warmup ratio of 0.06 is widely used in the literature and recommended. For BERT-NLI we recommend a higher learning rate of around 0.4.
- Note that these findings are only based on experiments with DeBERTaV3-base. Different variants of BERT such as RoBERTa might have different optimal hyperparameters and

smaller or larger versions of the same variant can also require different hyperparameters. As a rule of thumb, we recommend using the average hyperparameters recommended in the paper for the respective variant and model size. In our experience, this can lead to good performance without extensive hyperparameter search.

Table 36 – Best hyperparameters DeBERTaV3-base

dataset	sample	learning_rate	epochs	batch_size	hypothesis/context	learning_rate_importance	epochs_importance	batch_size_importance	hypothesis_importance
cap-sotu	5000	2.58923007964209E-05	80	16	template_not_nli_context	0.8	0.19	0.03	0.08
coronanet	5000	1.65733512483522E-05	70	16		0.68	0.03	0.25	
manifesto-8	5000	1.48701225826261E-05	100	16	template_not_nli_context	0.78	0.05	0.02	0.03
manifesto-military	3970	5.04252964208114E-06	100	16	template_not_nli_context	0.86	0.02	0.0	0.12
manifesto-morality	3188	7.01427868409888E-06	40	8	template_not_nli_context	0.87	0.03	0.0	0.03
manifesto-military	2500	1.05189875504116E-05	40	8	template_not_nli_context	0.87	0.0	0.02	0.11
sentiment-news-econ	2500	5.23582379342828E-06	70	8		0.95	0.08	0.02	
manifesto-morality	2500	7.38652140567614E-06	40	8	template_not_nli_context	0.86	0.0	0.05	0.06
cap-us-court	2500	3.92349836108803E-05	30	16		0.6	0.04	0.26	
cap-us-court	2500	3.92349836108803E-05	30	16		0.61	0.11	0.3	
cap-sotu	2500	1.40886242827464E-05	100	16	template_not_nli_context	0.84	0.15	0.0	0.06
coronanet	2500	1.47658491433165E-05	70	16		0.83	0.13	0.1	
manifesto-8	2500	2.88397427518993E-05	100	16	template_not_nli_context	0.75	0.14	0.03	0.06
manifesto-protectionism	2116	1.87661093918064E-05	80	8	template_not_nli_context	0.83	0.16	0.02	0.06
manifesto-military	1000	2.00024058043717E-05	100	16	template_not_nli_context	0.86	0.14	0.0	0.05
sentiment-news-econ	1000	1.56276552800989E-05	80	16		0.85	0.05	0.08	
manifesto-morality	1000	2.09418250238339E-05	80	8	template_not_nli_context	0.83	0.17	0.05	0.13
cap-us-court	1000	2.82461093113355E-05	80	16		0.67	0.11	0.21	
cap-sotu	1000	6.63962307985946E-06	90	16	template_not_nli_context	0.73	0.1	0.15	0.06

coronanet	1000	4.351278135123E-05	50	16		0.66	0.13	0.29	
manifesto-protectionism	1000	4.17607216178266E-05	40	8	template_not_nli_context	0.91	0.03	0.0	0.09
manifesto-8	1000	3.22345648097683E-05	100	16	template_not_nli	0.73	0.23	0.13	0.05
manifesto-military	500	1.18845539881688E-06	100	8	template_not_nli_context	0.91	0.08	0.0	0.12
sentiment-news-econ	500	2.79019028493244E-05	40	16		0.62	0.29	0.16	
manifesto-morality	500	2.11559334830169E-05	100	8	template_not_nli_context	0.75	0.09	0.01	0.02
cap-us-court	500	3.35265232133956E-05	50	16		0.6	0.14	0.33	
cap-sotu	500	0.00010277620057349700	100	32	template_not_nli_context	0.58	0.06	0.09	0.15
coronanet	500	4.3482519457373E-05	50	16		0.71	0.06	0.38	
manifesto-protectionism	500	6.93297654840942E-06	70	8	template_not_nli_context	0.86	0.09	0.06	0.01
manifesto-8	500	5.46212164028217E-05	100	32	template_not_nli	0.55	0.08	0.25	0.15
manifesto-military	100	2.00024058043717E-05	100	16	template_not_nli_context	0.49	0.2	0.07	0.26
sentiment-news-econ	100	2.60796565980958E-05	60	8		0.66	0.26	0.11	
manifesto-morality	100	6.62608564808309E-05	100	16	template_not_nli	0.57	0.28	0.03	0.06
cap-us-court	100	3.22595734118766E-05	40	16		0.32	0.1	0.6	
cap-sotu	100	1.79444232053641E-05	100	8	template_not_nli_context	0.1	0.53	0.25	0.11
coronanet	100	8.51571917827537E-05	70	16		0.38	0.09	0.51	
manifesto-protectionism	100	1.83490720490555E-06	60	8	template_not_nli_context	0.64	0.12	0.05	0.11
manifesto-8	100	1.05813350572485E-05	80	16	template_not_nli	0.2	0.64	0.08	0.11

Table 37 – Best hyperparameters DeBERTaV3-NLI

dataset	sample	learning_rate	epochs	batch_size	hypothesis/context	lr_warmup_ratio	learning_rate_importance	epochs_importance	batch_size_importance	hypothesis_importance	lr_warmup_ratio_importance
cap-sotu	5000	2.83282126163917E-05	35	32	template_quote_context	0.6	0.71	0.06	0.02	0.03	0.06
coronanet	5000	2.29795540619134E-05	35	16	template_quote_long_hypo	0.2	0.32	0.08	0.03	0.33	0.19
manifesto-8	5000	6.16501824659614E-05	5	16	template_quote_context_long_hypo	0.6	0.24	0.46	0.07	0.06	0.2
manifesto-military	3970	7.01478285944153E-06	25	8	template_quote_context	0.4	0.69	0.0	0.03	0.14	0.14
manifesto-morality	3188	1.15044268072156E-05	35	16	template_quote_2_context	0.4	0.56	0.03	0.1	0.07	0.25
manifesto-military	2500	8.41792392388798E-06	50	8	template_quote_context	0.4	0.58	0.0	0.07	0.18	0.19
sentiment-news-econ	2500	2.17138293595285E-06	5	8	template_quote	0.06	0.5	0.3	0.08	0.02	0.14
manifesto-morality	2500	5.22905501774918E-05	50	16	template_quote_context	0.06	0.75	0.05	0.0	0.05	0.17
cap-us-court	2500	1.78910146352391E-05	30	16	template_quote	0.4	0.05	0.11	0.07	0.01	0.73
cap-us-court	2500	1.78910146352391E-05	30	16	template_quote	0.4	0.03	0.14	0.06	0.04	0.71
cap-sotu	2500	9.29876274781711E-06	15	16	template_quote_context	0.4	0.25	0.22	0.0	0.31	0.16
coronanet	2500	5.88725746302143E-05	40	32	template_quote	0.6	0.21	0.56	0.01	0.07	0.24
manifesto-8	2500	6.56448160773369E-05	15	16	template_quote_context	0.06	0.71	0.06	0.02	0.08	0.27
manifesto-protectionism	2116	6.48948885125795E-06	35	8	template_quote_context	0.4	0.7	0.06	0.05	0.2	0.13

manifesto-military	1000	8.41792392388798E-06	40	16	template_quote_2_context	0.4	0.69	0.06	0.01	0.07	0.16
sentiment-news-econ	1000	2.19335631253761E-05	20	8	template_complex	0.06	0.57	0.25	0.05	0.01	0.06
manifesto-morality	1000	0.00011533459180491600	40	16	template_quote_2_context	0.6	0.69	0.03	0.0	0.05	0.16
cap-us-court	1000	3.27242966059094E-05	45	8	template_quote	0.6	0.26	0.04	0.21	0.05	0.34
cap-sotu	1000	2.18190599341266E-05	15	16	template_quote_context_long_hypo	0.2	0.19	0.13	0.28	0.3	0.05
coronanet	1000	5.78412687027874E-05	25	32	template_quote_long_hypo	0.06	0.38	0.27	0.05	0.15	0.13
manifesto-protectionism	1000	1.23647272498165E-05	15	16	template_quote_context	0.6	0.64	0.16	0.01	0.19	0.1
manifesto-8	1000	4.11937189631642E-05	10	16	template_quote_context_long_hypo	0.2	0.59	0.07	0.06	0.1	0.15
manifesto-military	500	5.90897570146346E-06	50	8	template_quote_2_context	0.4	0.78	0.0	0.0	0.11	0.17
sentiment-news-econ	500	1.83983652952772E-05	20	16	template_complex	0.2	0.67	0.15	0.04	0.11	0.03
manifesto-morality	500	2.71557707516919E-06	25	8	template_quote_context	0.06	0.65	0.12	0.05	0.16	0.11
cap-us-court	500	3.27242966059094E-05	45	8	template_quote	0.6	0.51	0.08	0.17	0.06	0.14
cap-sotu	500	3.99001407993242E-05	35	16	template_quote_context_long_hypo	0.2	0.12	0.19	0.03	0.71	0.07
coronanet	500	7.02626320544305E-05	50	32	template_quote_long_hypo	0.06	0.35	0.22	0.08	0.07	0.33
manifesto-protectionism	500	9.19880667845133E-05	40	16	template_quote_context	0.4	0.7	0.0	0.0	0.23	0.03
manifesto-8	500	4.66719989597539E-05	10	8	template_quote_context_long_hypo	0.2	0.21	0.22	0.02	0.32	0.23

manifesto-military	100	8.41792392388798E-06	15	8	template_quote_context	0.06	0.28	0.03	0.01	0.18	0.4
sentiment-news-econ	100	6.09208561627334E-06	25	16	template_complex	0.06	0.6	0.17	0.01	0.29	0.18
manifesto-morality	100	6.46422878923982E-05	45	8	template_quote_context	0.4	0.21	0.1	0.04	0.45	0.23
cap-us-court	100	1.41571701486137E-05	30	16	template_quote_long_hypo	0.2	0.07	0.37	0.22	0.18	0.06
cap-sotu	100	1.61349070389162E-06	50	32	template_quote_context	0.4	0.06	0.07	0.06	0.47	0.25
coronanet	100	3.1543990308331E-06	15	8	template_quote	0.06	0.08	0.17	0.05	0.47	0.03
manifesto-protectionism	100	0.00011300038015015400	40	8	template_quote	0.6	0.65	0.0	0.02	0.23	0.11
manifesto-8	100	0.00012141307774357400	30	32	template_quote_context	0.6	0.29	0.27	0.14	0.27	0.03

E4: Pre-processing and hyperparameters for classical algorithms

Our pre-processing for classical algorithms followed standard practice. For the TFIDF vectorizer, we removed stop words, used lower case and lemmatization. As part of the hyperparameter search, we also tested different n-gram ranges and removed words of varying maximum and minimum frequency. For the word vector input, we used the average vector of relevant words in the text. We used pre-trained GloVe word vectors (Pennington, Socher, and Manning 2014) with 300 dimensions trained on Common Crawl provided by the SpaCy library (en_core_web_lg-3.2.0, Montani et al. 2022). Prior unpublished work reported surprisingly inferior results for averaged word embeddings compared to character/word n-grams (Terechshenko et al. 2020). We assume that these preliminary results are due to sub-optimal pre-processing. Averaging word vectors has the disadvantage that there is no weight attributed to more or less important words, compared to TFIDF which does perform a form of weighting based on frequencies. To alleviate this limitation, we therefore discarded the vectors of less relevant words using part-of-speech-tagging and only included the vectors for the following parts-of-speech: ["NOUN", "ADJ", "VERB", "PROPN", "ADV", "INTJ", "PRON"].[19] Moreover, for those datasets where we could include the preceding and following sentences (see above), we applied twice the weight to the vectors from the target sentence before averaging the embeddings. This means that the classifier could attribute higher importance to the target sentence, while still receiving information from the context sentences.

The best hyperparameters used for each dataset and sample size are displayed in the tables below. As hyperparameter searches for classical algorithms are computationally cheap, we

[19] https://universaldependencies.org/u/pos/

do not discuss hyperparameter choices in detail and we only show the parameters for TFIDF vectorization, as they are very similar for classification with word embeddings.

Table 38 - Best hyperparameters SVM with TFIDF

dataset	sample	ngram	max_df	min_df	kernel	C	gamma	class_weight	coef0	degree	epochs	context
cap-sotu	10000	(1, 3)	0.8	0.01	poly	79.47	scale	balanced	2.01	7	5000	yes
coronanet	10000	(1, 3)	0.8	0.01	rbf	2.46	scale		2.05	27	5000	
manifesto-8	10000	(1, 3)	0.7	0.01	rbf	340.93	scale		7.71	48	5000	yes
cap-us-court	5426	(1, 3)	0.7	0.01	linear	1.19	scale	balanced	2.22	8	2000	
cap-us-court	5000	(1, 3)	0.7	0.01	linear	1.19	scale	balanced	2.22	8	2000	
cap-sotu	5000	(1, 3)	0.7	0.01	linear	2.57	auto	balanced	7.5	46	3000	yes
coronanet	5000	(1, 3)	0.8	0.01	rbf	7.45	scale		4.11	6	3000	
manifesto-8	5000	(1, 3)	0.8	0.01	poly	79.47	scale	balanced	2.01	7	5000	yes
manifesto-military	3970	(1, 3)	0.8	0.01	poly	79.47	scale	balanced	2.01	7	5000	yes
manifesto-morality	3188	(1, 3)	0.7	0.01	poly	311.74	scale		3.71	8	5000	yes
sentiment-news-econ	3000	(1, 3)	0.9	0.03	poly	18.61	scale		3.95	8	2000	
manifesto-military	2500	(1, 3)	0.8	0.01	linear	1.03	scale		22.89	42	3000	yes
sentiment-news-econ	2500	(1, 3)	0.7	0.03	rbf	28.66	scale		3.38	40	4000	
manifesto-morality	2500	(1, 3)	0.7	0.01	rbf	127.92	scale		7.63	12	2000	yes
cap-us-court	2500	(1, 3)	0.7	0.01	linear	1.19	scale	balanced	2.22	8	2000	
cap-sotu	2500	(1, 3)	0.7	0.01	linear	2.57	auto	balanced	7.5	46	3000	yes
coronanet	2500	(1, 3)	0.7	0.01	linear	2.47	auto		2.17	11	2000	
manifesto-8	2500	(1, 3)	0.8	0.01	linear	1.48	scale		23.78	29	7000	yes
manifesto-protectionism	2116	(1, 3)	0.8	0.01	rbf	4.52	scale		1.83	20	1000	yes
manifesto-military	1000	(1, 3)	0.7	0.01	linear	2.57	auto	balanced	7.5	46	3000	yes
sentiment-news-econ	1000	(1, 3)	0.8	0.03	rbf	5.67	scale		1.65	22	4000	
manifesto-morality	1000	(1, 3)	0.7	0.01	linear	2.57	auto	balanced	7.5	46	3000	yes
cap-us-court	1000	(1, 3)	0.7	0.01	linear	1.19	scale	balanced	2.22	8	2000	
cap-sotu	1000	(1, 3)	0.7	0.01	linear	2.57	auto	balanced	7.5	46	3000	yes
coronanet	1000	(1, 2)	0.9	0.01	poly	1.06	scale		16.95	1	5000	
manifesto-protectionism	1000	(1, 3)	0.7	0.03	rbf	3.0	scale		4.04	29	5000	yes
manifesto-8	1000	(1, 3)	0.7	0.01	linear	2.3	auto		1.45	40	1000	yes

manifesto-military	500	(1, 3)	0.7	0.01	linear	1.53	auto		40.16	34	2000	yes
sentiment-news-econ	500	(1, 3)	0.7	0.03	rbf	18.74	scale	balanced	8.66	3	1000	
manifesto-morality	500	(1, 3)	0.7	0.01	linear	2.57	auto	balanced	7.5	46	3000	yes
cap-us-court	500	(1, 3)	0.7	0.01	linear	1.19	scale	balanced	2.22	8	2000	
cap-sotu	500	(1, 2)	0.8	0.01	poly	520.17	auto		14.02	32	4000	yes
coronanet	500	(1, 2)	0.7	0.01	linear	15.47	auto	balanced	71.31	25	2000	
manifesto-protectionism	500	(1, 2)	0.9	0.03	rbf	38.81	scale		18.46	17	3000	yes
manifesto-8	500	(1, 3)	0.7	0.01	linear	2.57	auto	balanced	7.5	46	3000	yes
manifesto-military	100	(1, 2)	0.9	0.03	rbf	38.81	scale		18.46	17	3000	yes
sentiment-news-econ	100	(1, 2)	0.9	0.03	linear	2.03	scale		93.73	45	2000	
manifesto-morality	100	(1, 3)	0.7	0.03	linear	1.84	scale		3.49	42	3000	yes
cap-us-court	100	(1, 3)	0.9	0.06	poly	2.29	auto		25.47	19	7000	
cap-sotu	100	(1, 3)	0.7	0.03	linear	1.84	scale		3.49	42	3000	yes
coronanet	100	(1, 2)	0.9	0.03	linear	191.58	auto		7.16	2	1000	
manifesto-protectionism	100	(1, 3)	0.8	0.03	rbf	3.31	scale		1.23	19	5000	yes
manifesto-8	100	(1, 3)	0.8	0.03	rbf	111.4	auto		66.71	40	6000	no

Table 39 - Best hyperparameters logistic regression with TFIDF

dataset	sample	ngram	max_df	min_df	solver	C	class_weight	max_iter	warm_start	context
cap-sotu	10000	(1, 2)	0.9	0.01	liblinear	5.47		157	FALSE	yes
coronanet	10000	(1, 3)	0.8	0.01	liblinear	3.08		226	FALSE	
manifesto-8	10000	(1, 2)	0.9	0.01	sag	3.67		138	TRUE	yes
cap-us-court	5426	(1, 3)	0.9	0.01	liblinear	94.58	balanced	282	TRUE	
cap-us-court	5000	(1, 2)	0.7	0.01	lbfgs	415.4	balanced	872	FALSE	
cap-sotu	5000	(1, 2)	0.7	0.01	saga	112.09	balanced	495	FALSE	yes
coronanet	5000	(1, 2)	0.8	0.01	liblinear	5.04		327	FALSE	
manifesto-8	5000	(1, 2)	0.7	0.01	sag	3.82		140	FALSE	yes
manifesto-military	3970	(1, 2)	0.9	0.01	sag	4.46		52	TRUE	yes
manifesto-morality	3188	(1, 2)	0.7	0.01	sag	4.31		668	FALSE	yes
sentiment-news-econ	3000	(1, 3)	0.7	0.06	sag	546.5		735	TRUE	
manifesto-military	2500	(1, 2)	0.8	0.01	saga	5.03		87	TRUE	yes
sentiment-news-econ	2500	(1, 3)	0.8	0.06	sag	4.06		779	TRUE	

manifesto-morality	2500	(1, 2)	0.7	0.01	newton-cg	38.56		721	TRUE	yes
cap-us-court	2500	(1, 3)	0.9	0.01	liblinear	94.58	balanced	282	TRUE	
cap-sotu	2500	(1, 2)	0.7	0.01	saga	112.09	balanced	495	FALSE	yes
coronanet	2500	(1, 2)	0.8	0.01	saga	8.59		213	FALSE	
manifesto-8	2500	(1, 2)	0.7	0.01	sag	7.79		106	FALSE	yes
manifesto-protectionism	2116	(1, 2)	0.8	0.01	sag	3.67		750	TRUE	yes
manifesto-military	1000	(1, 2)	0.7	0.01	lbfgs	42.29		590	TRUE	yes
sentiment-news-econ	1000	(1, 2)	0.8	0.01	saga	8.59		217	FALSE	
manifesto-morality	1000	(1, 2)	0.7	0.01	sag	27.62		742	TRUE	yes
cap-us-court	1000	(1, 2)	0.7	0.01	saga	822.71	balanced	633	FALSE	
cap-sotu	1000	(1, 2)	0.7	0.01	saga	112.09	balanced	495	FALSE	yes
coronanet	1000	(1, 2)	0.8	0.01	liblinear	13.43		292	FALSE	
manifesto-protectionism	1000	(1, 2)	0.9	0.01	newton-cg	584.07		987	TRUE	yes
manifesto-8	1000	(1, 2)	0.7	0.01	saga	112.09	balanced	495	FALSE	yes
manifesto-military	500	(1, 2)	0.9	0.01	saga	786.01		103	TRUE	yes
sentiment-news-econ	500	(1, 3)	0.7	0.03	sag	198.65		731	TRUE	
manifesto-morality	500	(1, 2)	0.7	0.01	saga	112.09	balanced	495	FALSE	yes
cap-us-court	500	(1, 2)	0.7	0.01	lbfgs	415.4	balanced	872	FALSE	
cap-sotu	500	(1, 2)	0.7	0.01	saga	112.09	balanced	495	FALSE	yes
coronanet	500	(1, 2)	0.8	0.01	saga	772.47	balanced	825	FALSE	
manifesto-protectionism	500	(1, 2)	0.7	0.01	liblinear	209.71		363	FALSE	yes
manifesto-8	500	(1, 2)	0.8	0.01	sag	927.14		92	TRUE	yes
manifesto-military	100	(1, 3)	0.8	0.03	newton-cg	921.95		93	FALSE	yes
sentiment-news-econ	100	(1, 3)	0.7	0.03	sag	65.99		818	TRUE	
manifesto-morality	100	(1, 2)	0.8	0.01	saga	212.86		470	FALSE	yes
cap-us-court	100	(1, 3)	0.7	0.03	lbfgs	609.95	balanced	314	TRUE	
cap-sotu	100	(1, 2)	0.8	0.01	saga	413.21		374	TRUE	yes
coronanet	100	(1, 2)	0.8	0.03	liblinear	950.46	balanced	639	TRUE	
manifesto-protectionism	100	(1, 2)	0.7	0.01	saga	323.1		565	TRUE	yes
manifesto-8	100	(1, 3)	0.7	0.06	saga	965.67	balanced	142	TRUE	yes
cap-sotu	10000	(1, 2)	0.9	0.01	liblinear	5.47		157	FALSE	yes
coronanet	10000	(1, 3)	0.8	0.01	liblinear	3.08		226	FALSE	

Appendix F: Training time

Compute costs and training times are an important limitation of deep learning models. The table below displays the training time required for training a single algorithm with a given number of training examples averaged across our eight tasks. Classical algorithms are significantly faster on a CPU than BERT-like algorithms on high-performance GPUs. Note that, in practice, multiple algorithms need to be trained for hyperparameter search and calculating uncertainty and training time is therefore higher than simply training a single model.

At the same time, compute costs and hardware are much less of a hurdle than they were a few years ago. The analyses for this paper were initially set up in a Google Colab notebook, which provides easy access to GPUs in the browser. We used the 10 EUR / month subscription, which provides decent GPU run-times of theoretically up to 24 hours. In practice, we started our script described in appendix C and manually monitored our browser roughly every 30 minutes to make sure that the GPU run-time was not timed out due to inactivity. We tried to let the GPU run over night, which worked in around 50% of cases, while in 50% of cases Google had timed out our GPU. In our experience, this setup enabled GPU run-times between roughly 6 to 18 hours. To avoid losing data when the GPU timed out, we needed to add intermediate saving steps in our script. As we added more datasets and sample sizes, the random time outs of Google Colab became more and more inconvenient, and we switched to a university GPU. For users without access to university GPUs, newer Colab subscriptions promise more stable run-times for 50 EUR, but we have not tested how reliable they are.

Based on this experience, we learned that compute resources are an important hurdle for using deep learning, but it is less pronounced than we originally thought. Substantive research

projects do not need to compare many datasets across many data sizes, training hundreds of models, but only need to train a few dozen models for their specific dataset. Moreover, our extensive hyperparameter search described in appendix E shows, that the best performing hyperparameters always oscillate around a certain set of values. Researchers can probably save significant compute time if they chose default hyperparameters indicated in appendix E.

Table 40 - Training time comparison for a single model

algorithm	sample size	minutes training	hardware
SVM_tfidf	100.0	0.0	CPU (AMD Rome 7H12)
SVM_tfidf	500.0	0.0	CPU (AMD Rome 7H12)
SVM_tfidf	1000.0	0.0	CPU (AMD Rome 7H12)
SVM_tfidf	2500.0	0.0	CPU (AMD Rome 7H12)
SVM_tfidf	5000.0	0.5	CPU (AMD Rome 7H12)
SVM_tfidf	10000.0	1.0	CPU (AMD Rome 7H12)
logistic_tfidf	100.0	0.0	CPU (AMD Rome 7H12)
logistic_tfidf	500.0	0.0	CPU (AMD Rome 7H12)
logistic_tfidf	1000.0	0.0	CPU (AMD Rome 7H12)
logistic_tfidf	2500.0	0.0	CPU (AMD Rome 7H12)
logistic_tfidf	5000.0	0.0	CPU (AMD Rome 7H12)
logistic_tfidf	10000.0	0.0	CPU (AMD Rome 7H12)
SVM_embeddings	100.0	0.0	CPU (AMD Rome 7H12)
SVM_embeddings	500.0	0.0	CPU (AMD Rome 7H12)
SVM_embeddings	1000.0	0.0	CPU (AMD Rome 7H12)
SVM_embeddings	2500.0	0.0	CPU (AMD Rome 7H12)
SVM_embeddings	5000.0	0.0	CPU (AMD Rome 7H12)
SVM_embeddings	10000.0	0.67	CPU (AMD Rome 7H12)
logistic_embeddings	100.0	0.0	CPU (AMD Rome 7H12)
logistic_embeddings	500.0	0.0	CPU (AMD Rome 7H12)
logistic_embeddings	1000.0	0.0	CPU (AMD Rome 7H12)

logistic_embeddings	2500.0	0.0	CPU (AMD Rome 7H12)
logistic_embeddings	5000.0	0.0	CPU (AMD Rome 7H12)
logistic_embeddings	10000.0	0.33	CPU (AMD Rome 7H12)
BERT-base-nli	100.0	3.75	GPU (A100)
BERT-base-nli	500.0	8.75	GPU (A100)
BERT-base-nli	1000.0	11.62	GPU (A100)
BERT-base-nli	2500.0	23.43	GPU (A100)
BERT-base-nli	5000.0	44.0	GPU (A100)
BERT-base-nli	10000.0	45.0	GPU (A100)
BERT-base	100.0	1.38	GPU (A100)
BERT-base	500.0	5.5	GPU (A100)
BERT-base	1000.0	12.12	GPU (A100)
BERT-base	2500.0	24.57	GPU (A100)
BERT-base	5000.0	37.25	GPU (A100)
BERT-base	10000.0	67.33	GPU (A100)

Bibliography

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

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.

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.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. 'A Large Annotated Corpus for Learning Natural Language Inference'. *ArXiv:1508.05326 [Cs]*, August. http://arxiv.org/abs/1508.05326.

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.

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.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. 'Unsupervised Cross-Lingual Representation Learning at Scale'. *ArXiv:1911.02116 [Cs]*, April. http://arxiv.org/abs/1911.02116.

Conneau, Alexis, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. 'XNLI: Evaluating Cross-Lingual Sentence Representations'. *ArXiv:1809.05053 [Cs]*, September. http://arxiv.org/abs/1809.05053.

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.

Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. 'Annotation Artifacts in Natural Language Inference Data'. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 107–12. New Orleans, Louisiana: Association for Computational Linguistics. doi:10.18653/v1/N18-2017.

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.

Montani, Ines, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O'Leary McCann, et al. 2022. 'Explosion/SpaCy: V3.2.4'. Zenodo. doi:10.5281/ZENODO.6394862.

Nie, Yixin, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. 'Adversarial NLI: A New Benchmark for Natural Language Understanding'. *ArXiv:1910.14599 [Cs]*, May. http://arxiv.org/abs/1910.14599.

Parrish, Alicia, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alex Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. 'Does Putting a Linguist in the Loop Improve NLU Data Collection?' *ArXiv:2104.07179 [Cs]*, April. http://arxiv.org/abs/2104.07179.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. 'GloVe: Global Vectors for Word Representation'. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–43. Doha, Qatar: Association for Computational Linguistics. doi:10.3115/v1/D14-1162.

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.

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.

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.

Yin, Wenpeng, Dragomir Radev, and Caiming Xiong. 2021. 'DocNLI: A Large-Scale Dataset for Document-Level Natural Language Inference'. *ArXiv:2106.09449 [Cs]*, June. http://arxiv.org/abs/2106.09449.