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

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

Department of Communication Science, Vrije Universiteit Amsterdam, Netherlands

\*m.laurer@vu.nl

Online Appendix

[Appendix A: Dataset details 2](#_Toc108618812)

[Appendix B: Additional details on BERT and BERT-NLI 10](#_Toc108618813)

[Appendix C: Analysis Pipeline 20](#_Toc108618814)

[Appendix D: Metrics per algorithm per sample size 21](#_Toc108618815)

[Appendix E: Hyperparameters and Pre-processing 29](#_Toc108618816)

[Appendix F: Training time 43](#_Toc108618817)

[Bibliography 45](#_Toc108618818)

# 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 15 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]](#footnote-1)

**Manifesto Corpus (Burst et al. 2020)**

The Comparative Manifesto Project annotates party manifestos from political parties in over 50 countries since 1945.[[2]](#footnote-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 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]](#footnote-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 |

**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]](#footnote-4) The sentences are annotated based on 22 topical categories of the Comparative Agendas Project.[[5]](#footnote-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 |
| 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]](#footnote-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** |
| 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 |
| Transportation | 164 | 53 | 217 |
| Public Lands | 147 | 55 | 202 |
| Culture | 9 | 7 | 16 |

**CoronaNet (Cheng et al. 2020)**

The CoronaNet Research Project[[7]](#footnote-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 |
| 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 and BERT-NLI

**BERT-base**

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 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’).

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.).

**BERT-NLI**

Examples for Natural Language Inference (NLI) datasets are: SNLI with 570k examples (Bowman et al. 2015), MultiNLI with 433k (Williams, Nangia, and Bowman 2018), or ANLI with 162k and FEVER-NLI with 208k (Nie et al. 2020). These datasets cover domains like news articles, fictional literature, government documents, telephone conversations or image captions. 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).

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 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.

**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 {X}’.

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’.

**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** |
| 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 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 Administrative Configuration | 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.. |
| 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.[[8]](#footnote-8)

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[[9]](#footnote-9) for smart sampling of the best hyperparameters. We search over up to 60 hyperparameter configurations for the classical algorithm and up to 15 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]*

# Appendix D: Metrics per algorithm per sample size

The following tables display the exact metrics underlying the text and figures in the paper.

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 | 0,255 | 0,412 | 0,476 | 0,521 | 0,452 | 0,461 |
| **logistic\_tfidf** | 0 | 0,325 | 0,429 | 0,453 | 0,505 | 0,434 | 0,45 |
| **SVM\_embeddings** | 0 | 0,335 | 0,435 | 0,498 | 0,539 | 0,513 | 0,535 |
| **logistic\_embeddings** | 0 | 0,368 | 0,446 | 0,484 | 0,537 | 0,5 | 0,512 |
| **classical-best-tfidf** | 0 | 0,325 | 0,429 | 0,476 | 0,521 | 0,452 | 0,461 |
| **classical-best-embeddings** | 0 | 0,368 | 0,446 | 0,498 | 0,539 | 0,513 | 0,535 |
| **BERT-base** | 0 | 0,377 | 0,55 | 0,61 | 0,648 | 0,622 | 0,647 |
| **BERT-base-nli** | 0,403 | 0,486 | 0,591 | 0,62 | 0,647 | 0,601 | 0,63 |

Table 19 - Average F1-micro 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 | 0,447 | 0,62 | 0,654 | 0,691 | 0,571 | 0,565 |
| **logistic\_tfidf** | 0 | 0,539 | 0,623 | 0,637 | 0,683 | 0,549 | 0,557 |
| **SVM\_embeddings** | 0 | 0,552 | 0,637 | 0,662 | 0,691 | 0,607 | 0,604 |
| **logistic\_embeddings** | 0 | 0,566 | 0,626 | 0,655 | 0,701 | 0,596 | 0,585 |
| **classical-best-tfidf** | 0 | 0,539 | 0,623 | 0,654 | 0,691 | 0,571 | 0,565 |
| **classical-best-embeddings** | 0 | 0,566 | 0,637 | 0,662 | 0,701 | 0,607 | 0,604 |
| **BERT-base** | 0 | 0,584 | 0,71 | 0,746 | 0,769 | 0,686 | 0,693 |
| **BERT-base-nli** | 0,559 | 0,634 | 0,718 | 0,742 | 0,763 | 0,671 | 0,682 |

Table 20 - Average F1-macro performance difference between different algorithms

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Sample size** | **0 (8 datasets)** | **100 (8 datasets)** | **500 (8 datasets)** | **1000 (8 datasets)** | **2500 (8 datasets)** | **5000 (4 datasets)** | **10000 (3 datasets)** |
| **BERT-base vs. classical-best-tfidf** | 0 | 0,052 | 0,121 | 0,134 | 0,127 | 0,17 | 0,186 |
| **BERT-base vs. classical-best-embeddings** | 0 | 0,009 | 0,104 | 0,112 | 0,109 | 0,109 | 0,112 |
| **BERT-base-nli vs. classical-best-tfidf** | 0,403 | 0,161 | 0,162 | 0,144 | 0,126 | 0,149 | 0,169 |
| **BERT-base-nli vs. classical-best-embeddings** | 0,403 | 0,118 | 0,145 | 0,122 | 0,108 | 0,088 | 0,095 |
| **BERT-base-nli vs. BERT-base** | 0,403 | 0,109 | 0,041 | 0,01 | -0,001 | -0,021 | -0,017 |

Table 21 - Average F1-micro performance difference between different algorithms

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Sample size** | **0 (8 datasets)** | **100 (8 datasets)** | **500 (8 datasets)** | **1000 (8 datasets)** | **2500 (8 datasets)** | **5000 (4 datasets)** | **10000 (3 datasets)** |
| **BERT-base vs. classical-best-tfidf** | 0 | 0,045 | 0,087 | 0,092 | 0,078 | 0,115 | 0,128 |
| **BERT-base vs. classical-best-embeddings** | 0 | 0,018 | 0,073 | 0,084 | 0,068 | 0,079 | 0,089 |
| **BERT-base-nli vs. classical-best-tfidf** | 0,559 | 0,095 | 0,095 | 0,088 | 0,072 | 0,1 | 0,117 |
| **BERT-base-nli vs. classical-best-embeddings** | 0,559 | 0,068 | 0,081 | 0,08 | 0,062 | 0,064 | 0,078 |
| **BERT-base-nli vs. BERT-base** | 0,559 | 0,05 | 0,008 | -0,004 | -0,006 | -0,015 | -0,011 |

Table 22 - 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 | 0.0 | 0.0 | 0.02 |
| **f1\_macro\_mean** | 100 | 0.21 | 0.15 | 0.27 | 0.14 | 0.23 | 0.32 |
| **f1\_macro\_mean** | 500 | 0.29 | 0.3 | 0.34 | 0.35 | 0.27 | 0.45 |
| **f1\_macro\_mean** | 1000 | 0.3 | 0.3 | 0.37 | 0.37 | 0.43 | 0.47 |
| **f1\_macro\_mean** | 2500 | 0.34 | 0.35 | 0.4 | 0.41 | 0.48 | 0.49 |
| **f1\_macro\_mean** | 5000 | 0.34 | 0.38 | 0.41 | 0.43 | 0.5 | 0.51 |
| **f1\_macro\_mean** | 10000 | 0.39 | 0.4 | 0.42 | 0.44 | 0.52 | 0.53 |
| **f1\_micro\_mean** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.02 |
| **f1\_micro\_mean** | 100 | 0.36 | 0.34 | 0.38 | 0.36 | 0.35 | 0.38 |
| **f1\_micro\_mean** | 500 | 0.42 | 0.42 | 0.47 | 0.48 | 0.44 | 0.52 |
| **f1\_micro\_mean** | 1000 | 0.41 | 0.4 | 0.51 | 0.51 | 0.56 | 0.57 |
| **f1\_micro\_mean** | 2500 | 0.49 | 0.44 | 0.54 | 0.55 | 0.6 | 0.59 |
| **f1\_micro\_mean** | 5000 | 0.45 | 0.52 | 0.55 | 0.56 | 0.61 | 0.61 |
| **f1\_micro\_mean** | 10000 | 0.53 | 0.54 | 0.56 | 0.58 | 0.64 | 0.63 |
| **f1\_macro\_std** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_macro\_std** | 100 | 0.01 | 0.0 | 0.02 | 0.01 | 0.02 | 0.01 |
| **f1\_macro\_std** | 500 | 0.01 | 0.01 | 0.01 | 0.01 | 0.15 | 0.01 |
| **f1\_macro\_std** | 1000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 | 0.01 |
| **f1\_macro\_std** | 2500 | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 | 0.0 |
| **f1\_macro\_std** | 5000 | 0.0 | 0.0 | 0.01 | 0.01 | 0.0 | 0.01 |
| **f1\_macro\_std** | 10000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 | 0.0 |
| **f1\_micro\_std** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_micro\_std** | 100 | 0.01 | 0.01 | 0.02 | 0.02 | 0.01 | 0.03 |
| **f1\_micro\_std** | 500 | 0.0 | 0.01 | 0.0 | 0.0 | 0.09 | 0.01 |
| **f1\_micro\_std** | 1000 | 0.01 | 0.0 | 0.0 | 0.0 | 0.01 | 0.02 |
| **f1\_micro\_std** | 2500 | 0.0 | 0.01 | 0.0 | 0.0 | 0.0 | 0.01 |
| **f1\_micro\_std** | 5000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 | 0.0 |
| **f1\_micro\_std** | 10000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 |

Table 23 - 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 | 0.0 | 0.0 | 0.59 |
| **f1\_macro\_mean** | 100 | 0.4 | 0.26 | 0.48 | 0.46 | 0.59 | 0.59 |
| **f1\_macro\_mean** | 500 | 0.56 | 0.49 | 0.52 | 0.56 | 0.65 | 0.69 |
| **f1\_macro\_mean** | 1000 | 0.56 | 0.59 | 0.56 | 0.6 | 0.68 | 0.72 |
| **f1\_macro\_mean** | 2500 | 0.62 | 0.62 | 0.6 | 0.62 | 0.69 | 0.72 |
| **f1\_macro\_mean** | 3970 (all) | 0.63 | 0.67 | 0.6 | 0.64 | 0.72 | 0.75 |
| **f1\_micro\_mean** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.92 |
| **f1\_micro\_mean** | 100 | 0.77 | 0.42 | 0.78 | 0.76 | 0.9 | 0.88 |
| **f1\_micro\_mean** | 500 | 0.87 | 0.85 | 0.82 | 0.86 | 0.92 | 0.92 |
| **f1\_micro\_mean** | 1000 | 0.86 | 0.88 | 0.85 | 0.88 | 0.93 | 0.93 |
| **f1\_micro\_mean** | 2500 | 0.91 | 0.91 | 0.89 | 0.89 | 0.93 | 0.93 |
| **f1\_micro\_mean** | 3970 (all) | 0.91 | 0.93 | 0.88 | 0.9 | 0.94 | 0.94 |
| **f1\_macro\_std** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_macro\_std** | 100 | 0.02 | 0.01 | 0.02 | 0.01 | 0.02 | 0.04 |
| **f1\_macro\_std** | 500 | 0.01 | 0.0 | 0.02 | 0.02 | 0.03 | 0.01 |
| **f1\_macro\_std** | 1000 | 0.01 | 0.01 | 0.01 | 0.01 | 0.0 | 0.01 |
| **f1\_macro\_std** | 2500 | 0.01 | 0.01 | 0.01 | 0.0 | 0.02 | 0.01 |
| **f1\_macro\_std** | 3970 (all) | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 | 0.0 |
| **f1\_micro\_std** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_micro\_std** | 100 | 0.03 | 0.02 | 0.01 | 0.02 | 0.01 | 0.02 |
| **f1\_micro\_std** | 500 | 0.01 | 0.01 | 0.01 | 0.0 | 0.01 | 0.0 |
| **f1\_micro\_std** | 1000 | 0.01 | 0.01 | 0.01 | 0.01 | 0.0 | 0.0 |
| **f1\_micro\_std** | 2500 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_micro\_std** | 3970 (all) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

Table 24 - 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 | 0.0 | 0.0 | 0.38 |
| **f1\_macro\_mean** | 100 | 0.52 | 0.44 | 0.52 | 0.5 | 0.58 | 0.58 |
| **f1\_macro\_mean** | 500 | 0.53 | 0.55 | 0.6 | 0.54 | 0.64 | 0.69 |
| **f1\_macro\_mean** | 1000 | 0.5 | 0.64 | 0.55 | 0.61 | 0.69 | 0.7 |
| **f1\_macro\_mean** | 2116 (all) | 0.63 | 0.67 | 0.62 | 0.63 | 0.72 | 0.75 |
| **f1\_micro\_mean** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.72 |
| **f1\_micro\_mean** | 100 | 0.83 | 0.79 | 0.82 | 0.79 | 0.81 | 0.85 |
| **f1\_micro\_mean** | 500 | 0.83 | 0.87 | 0.87 | 0.82 | 0.86 | 0.9 |
| **f1\_micro\_mean** | 1000 | 0.79 | 0.91 | 0.82 | 0.87 | 0.9 | 0.9 |
| **f1\_micro\_mean** | 2116 (all) | 0.89 | 0.91 | 0.88 | 0.89 | 0.92 | 0.92 |
| **f1\_macro\_std** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_macro\_std** | 100 | 0.01 | 0.01 | 0.01 | 0.02 | 0.03 | 0.01 |
| **f1\_macro\_std** | 500 | 0.01 | 0.0 | 0.0 | 0.0 | 0.01 | 0.02 |
| **f1\_macro\_std** | 1000 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 |
| **f1\_macro\_std** | 2116 (all) | 0.0 | 0.0 | 0.0 | 0.0 | 0.02 | 0.02 |
| **f1\_micro\_std** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_micro\_std** | 100 | 0.02 | 0.02 | 0.02 | 0.04 | 0.05 | 0.01 |
| **f1\_micro\_std** | 500 | 0.01 | 0.01 | 0.0 | 0.0 | 0.02 | 0.01 |
| **f1\_micro\_std** | 1000 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| **f1\_micro\_std** | 2116 (all) | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 | 0.01 |

Table 25 - 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 | 0.0 | 0.0 | 0.46 |
| **f1\_macro\_mean** | 100 | 0.4 | 0.33 | 0.44 | 0.44 | 0.41 | 0.5 |
| **f1\_macro\_mean** | 500 | 0.44 | 0.46 | 0.42 | 0.49 | 0.59 | 0.6 |
| **f1\_macro\_mean** | 1000 | 0.49 | 0.5 | 0.49 | 0.55 | 0.67 | 0.64 |
| **f1\_macro\_mean** | 2500 | 0.53 | 0.57 | 0.56 | 0.56 | 0.71 | 0.7 |
| **f1\_macro\_mean** | 3188 (all) | 0.56 | 0.59 | 0.58 | 0.61 | 0.69 | 0.69 |
| **f1\_micro\_mean** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.9 |
| **f1\_micro\_mean** | 100 | 0.72 | 0.62 | 0.74 | 0.73 | 0.71 | 0.85 |
| **f1\_micro\_mean** | 500 | 0.75 | 0.78 | 0.7 | 0.79 | 0.86 | 0.87 |
| **f1\_micro\_mean** | 1000 | 0.8 | 0.82 | 0.78 | 0.84 | 0.9 | 0.88 |
| **f1\_micro\_mean** | 2500 | 0.85 | 0.89 | 0.84 | 0.83 | 0.91 | 0.91 |
| **f1\_micro\_mean** | 3188 (all) | 0.86 | 0.89 | 0.84 | 0.85 | 0.9 | 0.91 |
| **f1\_macro\_std** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_macro\_std** | 100 | 0.03 | 0.01 | 0.01 | 0.01 | 0.08 | 0.03 |
| **f1\_macro\_std** | 500 | 0.02 | 0.02 | 0.01 | 0.03 | 0.01 | 0.02 |
| **f1\_macro\_std** | 1000 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| **f1\_macro\_std** | 2500 | 0.0 | 0.01 | 0.01 | 0.0 | 0.01 | 0.02 |
| **f1\_macro\_std** | 3188 (all) | 0.0 | 0.0 | 0.0 | 0.0 | 0.02 | 0.02 |
| **f1\_micro\_std** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_micro\_std** | 100 | 0.06 | 0.03 | 0.01 | 0.02 | 0.13 | 0.02 |
| **f1\_micro\_std** | 500 | 0.02 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 |
| **f1\_micro\_std** | 1000 | 0.01 | 0.01 | 0.01 | 0.0 | 0.01 | 0.01 |
| **f1\_micro\_std** | 2500 | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 | 0.01 |
| **f1\_micro\_std** | 3188 (all) | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 | 0.0 |

Table 26 - 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 | 0.0 | 0.0 | 0.71 |
| **f1\_macro\_mean** | 100 | 0.56 | 0.55 | 0.58 | 0.57 | 0.56 | 0.67 |
| **f1\_macro\_mean** | 500 | 0.58 | 0.56 | 0.56 | 0.44 | 0.69 | 0.7 |
| **f1\_macro\_mean** | 1000 | 0.59 | 0.58 | 0.59 | 0.58 | 0.71 | 0.74 |
| **f1\_macro\_mean** | 2500 | 0.62 | 0.63 | 0.63 | 0.59 | 0.69 | 0.74 |
| **f1\_macro\_mean** | 3000 (all) | 0.63 | 0.64 | 0.62 | 0.58 | 0.7 | 0.74 |
| **f1\_micro\_mean** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.73 |
| **f1\_micro\_mean** | 100 | 0.64 | 0.64 | 0.61 | 0.63 | 0.64 | 0.69 |
| **f1\_micro\_mean** | 500 | 0.65 | 0.63 | 0.66 | 0.64 | 0.72 | 0.72 |
| **f1\_micro\_mean** | 1000 | 0.66 | 0.65 | 0.68 | 0.61 | 0.74 | 0.76 |
| **f1\_micro\_mean** | 2500 | 0.68 | 0.7 | 0.69 | 0.64 | 0.72 | 0.76 |
| **f1\_micro\_mean** | 3000 (all) | 0.69 | 0.69 | 0.68 | 0.66 | 0.73 | 0.75 |
| **f1\_macro\_std** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_macro\_std** | 100 | 0.06 | 0.07 | 0.03 | 0.04 | 0.1 | 0.02 |
| **f1\_macro\_std** | 500 | 0.01 | 0.03 | 0.05 | 0.04 | 0.02 | 0.01 |
| **f1\_macro\_std** | 1000 | 0.02 | 0.03 | 0.03 | 0.02 | 0.01 | 0.01 |
| **f1\_macro\_std** | 2500 | 0.02 | 0.01 | 0.01 | 0.0 | 0.02 | 0.01 |
| **f1\_macro\_std** | 3000 (all) | 0.0 | 0.0 | 0.0 | 0.01 | 0.01 | 0.0 |
| **f1\_micro\_std** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_micro\_std** | 100 | 0.02 | 0.02 | 0.04 | 0.01 | 0.05 | 0.02 |
| **f1\_micro\_std** | 500 | 0.02 | 0.02 | 0.02 | 0.0 | 0.02 | 0.01 |
| **f1\_micro\_std** | 1000 | 0.01 | 0.02 | 0.01 | 0.02 | 0.0 | 0.01 |
| **f1\_micro\_std** | 2500 | 0.02 | 0.01 | 0.01 | 0.0 | 0.01 | 0.01 |
| **f1\_micro\_std** | 3000 (all) | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 | 0.0 |

Table 27 - 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 | 0.0 | 0.0 | 0.36 |
| **f1\_macro\_mean** | 100 | 0.12 | 0.11 | 0.16 | 0.14 | 0.2 | 0.47 |
| **f1\_macro\_mean** | 500 | 0.29 | 0.28 | 0.3 | 0.29 | 0.47 | 0.52 |
| **f1\_macro\_mean** | 1000 | 0.34 | 0.35 | 0.37 | 0.36 | 0.52 | 0.55 |
| **f1\_macro\_mean** | 2500 | 0.36 | 0.38 | 0.42 | 0.43 | 0.62 | 0.59 |
| **f1\_macro\_mean** | 5000 | 0.39 | 0.41 | 0.46 | 0.48 | 0.68 | 0.63 |
| **f1\_macro\_mean** | 10000 | 0.41 | 0.42 | 0.49 | 0.52 | 0.68 | 0.66 |
| **f1\_micro\_mean** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.42 |
| **f1\_micro\_mean** | 100 | 0.25 | 0.23 | 0.34 | 0.32 | 0.43 | 0.54 |
| **f1\_micro\_mean** | 500 | 0.43 | 0.42 | 0.41 | 0.43 | 0.59 | 0.57 |
| **f1\_micro\_mean** | 1000 | 0.46 | 0.48 | 0.47 | 0.45 | 0.6 | 0.59 |
| **f1\_micro\_mean** | 2500 | 0.47 | 0.5 | 0.51 | 0.49 | 0.67 | 0.64 |
| **f1\_micro\_mean** | 5000 | 0.52 | 0.53 | 0.54 | 0.54 | 0.7 | 0.67 |
| **f1\_micro\_mean** | 10000 | 0.54 | 0.55 | 0.56 | 0.57 | 0.7 | 0.7 |
| **f1\_macro\_std** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_macro\_std** | 100 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.01 |
| **f1\_macro\_std** | 500 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.01 |
| **f1\_macro\_std** | 1000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.02 | 0.01 |
| **f1\_macro\_std** | 2500 | 0.0 | 0.01 | 0.01 | 0.01 | 0.02 | 0.01 |
| **f1\_macro\_std** | 5000 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| **f1\_macro\_std** | 10000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.02 |
| **f1\_micro\_std** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_micro\_std** | 100 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 |
| **f1\_micro\_std** | 500 | 0.01 | 0.01 | 0.0 | 0.01 | 0.01 | 0.02 |
| **f1\_micro\_std** | 1000 | 0.01 | 0.01 | 0.0 | 0.0 | 0.01 | 0.02 |
| **f1\_micro\_std** | 2500 | 0.0 | 0.0 | 0.01 | 0.0 | 0.0 | 0.01 |
| **f1\_micro\_std** | 5000 | 0.0 | 0.0 | 0.0 | 0.01 | 0.01 | 0.01 |
| **f1\_micro\_std** | 10000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 |

Table 28 - 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 | 0.0 | 0.0 | 0.32 |
| **f1\_macro\_mean** | 100 | 0.17 | 0.14 | 0.24 | 0.21 | 0.16 | 0.28 |
| **f1\_macro\_mean** | 500 | 0.34 | 0.27 | 0.39 | 0.35 | 0.49 | 0.51 |
| **f1\_macro\_mean** | 1000 | 0.41 | 0.41 | 0.45 | 0.43 | 0.52 | 0.54 |
| **f1\_macro\_mean** | 2500 | 0.45 | 0.46 | 0.5 | 0.5 | 0.57 | 0.55 |
| **f1\_macro\_mean** | 5000 | 0.48 | 0.5 | 0.53 | 0.54 | 0.6 | 0.58 |
| **f1\_macro\_mean** | 5426 (all) | 0.49 | 0.51 | 0.54 | 0.55 | 0.6 | 0.57 |
| **f1\_micro\_mean** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.39 |
| **f1\_micro\_mean** | 100 | 0.45 | 0.44 | 0.51 | 0.5 | 0.44 | 0.4 |
| **f1\_micro\_mean** | 500 | 0.59 | 0.56 | 0.58 | 0.59 | 0.65 | 0.65 |
| **f1\_micro\_mean** | 1000 | 0.63 | 0.62 | 0.6 | 0.62 | 0.67 | 0.67 |
| **f1\_micro\_mean** | 2500 | 0.64 | 0.65 | 0.67 | 0.64 | 0.7 | 0.68 |
| **f1\_micro\_mean** | 5000 | 0.65 | 0.67 | 0.67 | 0.69 | 0.71 | 0.71 |
| **f1\_micro\_mean** | 5426 (all) | 0.66 | 0.67 | 0.68 | 0.69 | 0.72 | 0.7 |
| **f1\_macro\_std** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_macro\_std** | 100 | 0.02 | 0.01 | 0.02 | 0.03 | 0.1 | 0.13 |
| **f1\_macro\_std** | 500 | 0.03 | 0.03 | 0.01 | 0.01 | 0.02 | 0.01 |
| **f1\_macro\_std** | 1000 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| **f1\_macro\_std** | 2500 | 0.0 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| **f1\_macro\_std** | 5000 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.0 |
| **f1\_macro\_std** | 5426 (all) | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 | 0.01 |
| **f1\_micro\_std** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_micro\_std** | 100 | 0.0 | 0.01 | 0.01 | 0.01 | 0.09 | 0.16 |
| **f1\_micro\_std** | 500 | 0.02 | 0.02 | 0.0 | 0.01 | 0.01 | 0.01 |
| **f1\_micro\_std** | 1000 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| **f1\_micro\_std** | 2500 | 0.0 | 0.0 | 0.0 | 0.01 | 0.0 | 0.01 |
| **f1\_micro\_std** | 5000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 | 0.01 |
| **f1\_micro\_std** | 5426 (all) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 |

Table 29 - 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 | 0.0 | 0.0 | 0.39 |
| **f1\_macro\_mean** | 100 | 0.22 | 0.04 | 0.25 | 0.22 | 0.29 | 0.47 |
| **f1\_macro\_mean** | 500 | 0.39 | 0.38 | 0.44 | 0.45 | 0.6 | 0.57 |
| **f1\_macro\_mean** | 1000 | 0.43 | 0.45 | 0.5 | 0.48 | 0.65 | 0.61 |
| **f1\_macro\_mean** | 2500 | 0.49 | 0.49 | 0.56 | 0.57 | 0.69 | 0.64 |
| **f1\_macro\_mean** | 5000 | 0.52 | 0.53 | 0.6 | 0.61 | 0.71 | 0.68 |
| **f1\_macro\_mean** | 10000 | 0.55 | 0.57 | 0.62 | 0.64 | 0.74 | 0.7 |
| **f1\_micro\_mean** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.37 |
| **f1\_micro\_mean** | 100 | 0.29 | 0.09 | 0.34 | 0.32 | 0.38 | 0.48 |
| **f1\_micro\_mean** | 500 | 0.44 | 0.43 | 0.49 | 0.48 | 0.62 | 0.59 |
| **f1\_micro\_mean** | 1000 | 0.47 | 0.48 | 0.52 | 0.52 | 0.67 | 0.63 |
| **f1\_micro\_mean** | 2500 | 0.54 | 0.51 | 0.6 | 0.6 | 0.7 | 0.66 |
| **f1\_micro\_mean** | 5000 | 0.57 | 0.57 | 0.62 | 0.63 | 0.72 | 0.7 |
| **f1\_micro\_mean** | 10000 | 0.6 | 0.61 | 0.64 | 0.66 | 0.74 | 0.72 |
| **f1\_macro\_std** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_macro\_std** | 100 | 0.0 | 0.03 | 0.02 | 0.01 | 0.03 | 0.02 |
| **f1\_macro\_std** | 500 | 0.01 | 0.01 | 0.01 | 0.0 | 0.05 | 0.02 |
| **f1\_macro\_std** | 1000 | 0.01 | 0.0 | 0.0 | 0.0 | 0.01 | 0.02 |
| **f1\_macro\_std** | 2500 | 0.01 | 0.01 | 0.0 | 0.0 | 0.01 | 0.01 |
| **f1\_macro\_std** | 5000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 | 0.01 |
| **f1\_macro\_std** | 10000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 |
| **f1\_micro\_std** | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1\_micro\_std** | 100 | 0.02 | 0.02 | 0.03 | 0.03 | 0.02 | 0.01 |
| **f1\_micro\_std** | 500 | 0.01 | 0.01 | 0.01 | 0.01 | 0.04 | 0.02 |
| **f1\_micro\_std** | 1000 | 0.01 | 0.0 | 0.0 | 0.0 | 0.01 | 0.01 |
| **f1\_micro\_std** | 2500 | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 | 0.01 |
| **f1\_micro\_std** | 5000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 | 0.0 |
| **f1\_micro\_std** | 10000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.01 |

Figure 1 - F1-micro performance per dataset per sample size per algorithm

**A picture containing diagram

Description automatically generated**

# Appendix E: Hyperparameters and Pre-processing

The tables below display the best hyperparameters for each dataset and sample size determined by a hyperparameter search with the Python library Optuna[[10]](#footnote-10) 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 and for BERT models up to 15 configurations were tested to save computational resources. Moreover, for BERT models no hyperparameter search was conducted for sample size 10 000. 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).[[11]](#footnote-11)

**Including context sentences or not?**

For quasi-sentence level datasets (Manifesto datasets and State of the Union Speeches), we also tested concatenating the target sentence with its preceding and following 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.[[12]](#footnote-12)

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 and instead of isolated strings. 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. This method 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 implemented the train-test-split for these datasets on the document level instead of the sentence level.

Table 30 - Likelihood of surrounding sentences having the same label as the target sentence

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

**Choosing hyperparameters - advice for BERT algorithms**

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 warm up, 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. 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 36 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, less epochs also lead to good performance. For datasets larger than 10000, less than 10 epochs can be used.
* 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-7, 9e-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.
* 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 31 - Best hyperparameters DeBERTa-base

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **dataset** | **sample** | **learning\_rate** | **epochs** | **batch\_size** | **context** | **learning\_rate\_importance** | **epochs\_importance** | **batch\_size\_importance** | **context\_importance** |
| coronanet | 5000 | 9.47351426808452e-06 | 70 | 16 |  | 0.71 | 0.11 | 0.23 |  |
| manifesto-8 | 5000 | 1.736918415611196e-05 | 100 | 16 | template\_not\_nli\_context | 0.81 | 0.11 | 0.01 | 0.08 |
| cap-sotu | 5000 | 2.589230079642088e-05 | 80 | 16 | template\_not\_nli\_context | 0.75 | 0.06 | 0.05 | 0.08 |
| manifesto-military | 3970 | 6.9660208852401204e-06 | 100 | 16 | template\_not\_nli\_context | 0.53 | 0.06 | 0.19 | 0.33 |
| manifesto-morality | 3188 | 1.7494010419777636e-05 | 90 | 16 | template\_not\_nli\_context | 0.5 | 0.06 | 0.18 | 0.22 |
| manifesto-military | 2500 | 6.566919215299559e-06 | 30 | 8 | template\_not\_nli\_context | 0.84 | 0.0 | 0.0 | 0.09 |
| coronanet | 2500 | 1.736918415611196e-05 | 20 | 16 |  | 0.58 | 0.1 | 0.29 |  |
| manifesto-8 | 2500 | 1.215176802593601e-05 | 100 | 16 | template\_not\_nli\_context | 0.87 | 0.08 | 0.0 | 0.07 |
| cap-us-court | 2500 | 2.5538488394915197e-05 | 30 | 16 |  | 0.69 | 0.08 | 0.27 |  |
| manifesto-morality | 2500 | 1.2969344702779084e-05 | 100 | 8 | template\_not\_nli\_context | 0.73 | 0.11 | 0.05 | 0.05 |
| sentiment-news-econ | 2500 | 1.2859734262104187e-05 | 40 | 8 |  | 0.94 | 0.06 | 0.01 |  |
| cap-sotu | 2500 | 2.241853457420366e-05 | 100 | 16 | template\_not\_nli\_context | 0.82 | 0.17 | 0.0 | 0.08 |
| manifesto-protectionism | 2116 | 2.1309243844008938e-05 | 80 | 16 | template\_not\_nli\_context | 0.86 | 0.16 | 0.03 | 0.08 |
| manifesto-military | 1000 | 7.295686093122079e-06 | 40 | 16 | template\_not\_nli\_context | 0.47 | 0.06 | 0.17 | 0.37 |
| coronanet | 1000 | 2.9031714046347617e-05 | 20 | 16 |  | 0.71 | 0.01 | 0.23 |  |
| manifesto-8 | 1000 | 7.941507417113066e-06 | 90 | 16 | template\_not\_nli\_context | 0.66 | 0.17 | 0.18 | 0.05 |
| cap-us-court | 1000 | 3.9073837979740804e-05 | 80 | 16 |  | 0.67 | 0.08 | 0.24 |  |
| manifesto-morality | 1000 | 7.941507417113066e-06 | 90 | 8 | template\_not\_nli\_context | 0.65 | 0.13 | 0.18 | 0.0 |
| sentiment-news-econ | 1000 | 9.14518166313462e-06 | 90 | 8 |  | 0.77 | 0.19 | 0.08 |  |
| manifesto-protectionism | 1000 | 7.295686093122079e-06 | 40 | 16 | template\_not\_nli\_context | 0.51 | 0.02 | 0.14 | 0.33 |
| cap-sotu | 1000 | 7.941507417113066e-06 | 90 | 16 | template\_not\_nli\_context | 0.53 | 0.12 | 0.19 | 0.04 |
| manifesto-military | 500 | 5.159999077362396e-06 | 50 | 8 | template\_not\_nli\_context | 0.67 | 0.03 | 0.05 | 0.19 |
| coronanet | 500 | 5.8398456254184985e-05 | 40 | 16 |  | 0.6 | 0.04 | 0.37 |  |
| manifesto-8 | 500 | 7.97413401808571e-05 | 100 | 32 | template\_not\_nli | 0.51 | 0.13 | 0.09 | 0.11 |
| cap-us-court | 500 | 2.9031714046347617e-05 | 40 | 16 |  | 0.78 | 0.12 | 0.05 |  |
| manifesto-morality | 500 | 7.941507417113066e-06 | 90 | 8 | template\_not\_nli\_context | 0.58 | 0.09 | 0.36 | 0.07 |
| manifesto-morality | 500 | 7.941507417113066e-06 | 90 | 8 | template\_not\_nli\_context | 0.49 | 0.09 | 0.33 | 0.03 |
| sentiment-news-econ | 500 | 8.898805178683114e-06 | 100 | 8 |  | 0.81 | 0.03 | 0.09 |  |
| manifesto-protectionism | 500 | 7.295686093122079e-06 | 40 | 16 | template\_not\_nli\_context | 0.4 | 0.16 | 0.17 | 0.22 |
| cap-sotu | 500 | 1.9568482000737615e-05 | 70 | 32 | template\_not\_nli\_context | 0.68 | 0.1 | 0.13 | 0.13 |
| manifesto-military | 100 | 7.941507417113066e-06 | 90 | 8 | template\_not\_nli\_context | 0.46 | 0.05 | 0.25 | 0.28 |
| coronanet | 100 | 6.217268263665875e-05 | 70 | 16 |  | 0.46 | 0.06 | 0.42 |  |
| manifesto-8 | 100 | 2.6965062743136563e-05 | 80 | 32 | template\_not\_nli | 0.28 | 0.32 | 0.27 | 0.15 |
| cap-us-court | 100 | 9.562386542636373e-05 | 60 | 16 |  | 0.19 | 0.11 | 0.65 |  |
| manifesto-morality | 100 | 5.9681471257977086e-05 | 100 | 16 | template\_not\_nli | 0.6 | 0.13 | 0.18 | 0.06 |
| manifesto-morality | 100 | 5.9681471257977086e-05 | 100 | 16 | template\_not\_nli | 0.6 | 0.15 | 0.16 | 0.02 |
| sentiment-news-econ | 100 | 3.889419392746646e-06 | 90 | 8 |  | 0.05 | 0.17 | 0.68 |  |
| manifesto-protectionism | 100 | 2.378807042347486e-05 | 60 | 16 | template\_not\_nli\_context | 0.48 | 0.06 | 0.11 | 0.39 |
| cap-sotu | 100 | 2.357914090850661e-05 | 100 | 8 | template\_not\_nli\_context | 0.21 | 0.55 | 0.11 | 0.14 |

Table 32 - Best hyperparameters DeBERTa-NLI

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **dataset** | **sample** | **learning\_rate** | **epochs** | **batch\_size** | **hypothesis/context** | **lr\_warmup\_ratio[[13]](#footnote-13)** | **learning\_rate\_importance** | **epochs\_importance** | **batch\_size\_importance** | **hypothesis\_importance** | **lr\_warmup\_ratio\_importance** |
| coronanet | 5000 | 4.6197903071938266e-05 | 31 | 32 | template\_quote\_long\_hypo | 0.25 | 0.17 | 0.5 | 0.03 | 0.19 | 0.04 |
| manifesto-8 | 5000 | 2.5610954358265684e-05 | 23 | 32 | template\_quote\_context\_long\_hypo | 0.25 | 0.03 | 0.07 | 0.01 | 0.46 | 0.41 |
| cap-sotu | 5000 | 2.910782974661089e-05 | 23 | 32 | template\_quote\_context\_long\_hypo | 0.25 | 0.05 | 0.62 | 0.01 | 0.13 | 0.08 |
| manifesto-military | 3970 | 2.08590685988187e-05 | 33 | 16 | template\_quote\_context | 0.5 | 0.57 | 0.05 | 0.17 | 0.19 | 0.05 |
| manifesto-morality | 3188 | 5.139069476807434e-05 | 25 | 8 | template\_quote\_context | 0.5 | 0.58 | 0.12 | 0.04 | 0.32 | 0.07 |
| manifesto-military | 2500 | 5.3629570679671993e-05 | 19 | 16 | template\_quote\_context |  | 0.53 | 0.16 | 0.13 | 0.05 |  |
| coronanet | 2500 | 1.2217032646463444e-05 | 7 | 32 | template\_quote |  | 0.67 | 0.03 | 0.06 | 0.33 |  |
| manifesto-8 | 2500 | 8.314470450162983e-05 | 17 | 32 | template\_quote\_context\_long\_hypo |  | 0.63 | 0.2 | 0.19 | 0.04 |  |
| cap-us-court | 2500 | 4.6591989032151934e-05 | 3 | 32 | template\_quote |  | 0.77 | 0.04 | 0.06 | 0.18 |  |
| manifesto-morality | 2500 | 1.7219395428516444e-05 | 17 | 16 | template\_quote\_2\_context |  | 0.45 | 0.17 | 0.16 | 0.02 |  |
| sentiment-news-econ | 2500 | 1.4177205879046634e-06 | 11 | 8 | template\_complex |  | 0.93 | 0.03 | 0.03 | 0.05 |  |
| cap-sotu | 2500 | 5.3629570679671993e-05 | 19 | 32 | template\_quote\_context |  | 0.82 | 0.12 | 0.13 | 0.04 |  |
| manifesto-protectionism | 2116 | 2.2174357641196533e-05 | 17 | 16 | template\_quote\_2\_context |  | 0.5 | 0.15 | 0.17 | 0.03 |  |
| manifesto-military | 1000 | 5.469973624651568e-06 | 9 | 8 | template\_quote\_context |  | 0.61 | 0.12 | 0.01 | 0.22 |  |
| coronanet | 1000 | 6.675214389155407e-05 | 11 | 32 | template\_quote |  | 0.1 | 0.46 | 0.35 | 0.17 |  |
| manifesto-8 | 1000 | 2.200649469244444e-05 | 9 | 16 | template\_quote\_context |  | 0.1 | 0.15 | 0.68 | 0.05 |  |
| cap-us-court | 1000 | 3.253779180787796e-05 | 13 | 32 | template\_quote |  | 0.14 | 0.41 | 0.18 | 0.19 |  |
| manifesto-morality | 1000 | 2.200649469244444e-05 | 9 | 8 | template\_quote\_context |  | 0.48 | 0.36 | 0.06 | 0.14 |  |
| sentiment-news-econ | 1000 | 1.601462283869358e-06 | 13 | 8 | template\_complex |  | 0.62 | 0.19 | 0.14 | 0.05 |  |
| manifesto-protectionism | 1000 | 2.200649469244444e-05 | 9 | 8 | template\_quote\_context |  | 0.74 | 0.16 | 0.14 | 0.09 |  |
| cap-sotu | 1000 | 1.2217032646463444e-05 | 9 | 16 | template\_quote\_context |  | 0.48 | 0.19 | 0.01 | 0.26 |  |
| manifesto-military | 500 | 2.200649469244444e-05 | 7 | 8 | template\_quote\_context |  | 0.74 | 0.02 | 0.08 | 0.01 |  |
| coronanet | 500 | 3.5107205953493424e-05 | 11 | 32 | template\_quote |  | 0.1 | 0.62 | 0.07 | 0.17 |  |
| manifesto-8 | 500 | 2.200649469244444e-05 | 7 | 16 | template\_quote\_context |  | 0.49 | 0.28 | 0.17 | 0.14 |  |
| cap-us-court | 500 | 6.675214389155407e-05 | 9 | 32 | template\_quote |  | 0.05 | 0.63 | 0.11 | 0.12 |  |
| manifesto-morality | 500 | 1.2217032646463444e-05 | 9 | 8 | template\_quote\_context |  | 0.6 | 0.17 | 0.01 | 0.12 |  |
| sentiment-news-econ | 500 | 5.258168640171172e-06 | 9 | 8 | template\_complex |  | 0.56 | 0.11 | 0.22 | 0.16 |  |
| sentiment-news-econ | 500 | 1.601462283869358e-06 | 15 | 8 | template\_complex |  | 0.5 | 0.12 | 0.19 | 0.13 |  |
| manifesto-protectionism | 500 | 2.200649469244444e-05 | 7 | 8 | template\_quote\_context |  | 0.75 | 0.18 | 0.21 | 0.07 |  |
| cap-sotu | 500 | 2.200649469244444e-05 | 7 | 16 | template\_quote\_context |  | 0.71 | 0.1 | 0.02 | 0.12 |  |
| manifesto-military | 100 | 4.4725914938671344e-06 | 9 | 8 | template\_quote\_2 |  | 0.78 | 0.09 | 0.04 | 0.04 |  |
| coronanet | 100 | 1.941451498370416e-05 | 11 | 32 | template\_quote |  | 0.08 | 0.38 | 0.27 | 0.28 |  |
| manifesto-8 | 100 | 6.438997796729219e-05 | 11 | 16 | template\_quote |  | 0.25 | 0.25 | 0.18 | 0.25 |  |
| cap-us-court | 100 | 8.190329427496868e-05 | 9 | 32 | template\_quote |  | 0.19 | 0.36 | 0.35 | 0.25 |  |
| manifesto-morality | 100 | 2.513841230249365e-05 | 15 | 8 | template\_quote |  | 0.49 | 0.16 | 0.12 | 0.1 |  |
| sentiment-news-econ | 100 | 2.0871100378973956e-05 | 7 | 4 | template\_quote |  | 0.84 | 0.09 | 0.09 | 0.0 |  |
| sentiment-news-econ | 100 | 1.242790511988806e-05 | 17 | 16 | template\_quote |  | 0.63 | 0.11 | 0.17 | 0.01 |  |
| manifesto-protectionism | 100 | 3.715821709222984e-05 | 11 | 8 | template\_quote |  | 0.56 | 0.11 | 0.15 | 0.03 |  |
| cap-sotu | 100 | 1.2217032646463444e-05 | 11 | 16 | template\_quote\_context |  | 0.66 | 0.11 | 0.19 | 0.21 |  |

**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. Averaging word vectors has the disadvantage that there is no weight attributed to more or less important words. 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"].[[14]](#footnote-14) 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 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 33 - Best hyperparameters SVM with TFIDF

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **dataset** | **sample** | **ngram** | **max\_df** | **min\_df** | **kernel** | **C** | **gamma** | **class\_weight** | **coef0** | **degree** | **epochs** | **context** |
| coronanet | 10000 | (1, 3) | 0.8 | 0.01 | rbf | 129.7 | scale |  | 90.51 | 17 | 5000 |  |
| manifesto-8 | 10000 | (1, 3) | 0.8 | 0.01 | poly | 79.47 | scale | balanced | 2.01 | 7 | 5000 | yes |
| cap-sotu | 10000 | (1, 3) | 0.8 | 0.01 | poly | 79.47 | scale | balanced | 2.01 | 7 | 5000 | yes |
| cap-us-court | 5426 | (1, 3) | 0.7 | 0.01 | linear | 1.19 | scale | balanced | 2.22 | 8 | 2000 |  |
| coronanet | 5000 | (1, 2) | 0.8 | 0.01 | rbf | 9.43 | scale |  | 2.88 | 11 | 3000 |  |
| manifesto-8 | 5000 | (1, 3) | 0.7 | 0.01 | poly | 26.35 | scale |  | 1.59 | 5 | 5000 | yes |
| 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 | poly | 180.35 | auto |  | 3.63 | 2 | 3000 | 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 | rbf | 155.84 | scale |  | 2.67 | 13 | 3000 | yes |
| sentiment-news-econ | 3000 | (1, 3) | 0.9 | 0.03 | rbf | 9.51 | scale |  | 78.54 | 5 | 3000 |  |
| manifesto-military | 2500 | (1, 3) | 0.8 | 0.01 | poly | 343.51 | scale |  | 2.67 | 4 | 3000 | yes |
| coronanet | 2500 | (1, 3) | 0.7 | 0.01 | linear | 1.19 | scale | balanced | 2.22 | 8 | 2000 |  |
| manifesto-8 | 2500 | (1, 3) | 0.7 | 0.01 | linear | 2.57 | auto | balanced | 7.5 | 46 | 3000 | yes |
| cap-us-court | 2500 | (1, 3) | 0.7 | 0.01 | linear | 1.19 | scale | balanced | 2.22 | 8 | 2000 |  |
| manifesto-morality | 2500 | (1, 3) | 0.8 | 0.01 | poly | 79.47 | scale | balanced | 2.01 | 7 | 5000 | yes |
| sentiment-news-econ | 2500 | (1, 3) | 0.9 | 0.03 | rbf | 1.99 | scale |  | 56.51 | 27 | 2000 |  |
| cap-sotu | 2500 | (1, 3) | 0.7 | 0.01 | linear | 2.57 | auto | balanced | 7.5 | 46 | 3000 | yes |
| manifesto-protectionism | 2116 | (1, 3) | 0.7 | 0.01 | rbf | 413.43 | scale |  | 32.78 | 14 | 2000 | yes |
| manifesto-military | 1000 | (1, 3) | 0.7 | 0.01 | linear | 2.57 | auto | balanced | 7.5 | 46 | 3000 | yes |
| coronanet | 1000 | (1, 3) | 0.7 | 0.01 | linear | 1.19 | scale | balanced | 2.22 | 8 | 2000 |  |
| manifesto-8 | 1000 | (1, 3) | 0.7 | 0.01 | poly | 189.55 | auto |  | 5.3 | 47 | 2000 | yes |
| cap-us-court | 1000 | (1, 3) | 0.7 | 0.01 | linear | 1.19 | scale | balanced | 2.22 | 8 | 2000 |  |
| manifesto-morality | 1000 | (1, 3) | 0.7 | 0.01 | linear | 2.57 | auto | balanced | 7.5 | 46 | 3000 | yes |
| sentiment-news-econ | 1000 | (1, 3) | 0.9 | 0.03 | rbf | 4.35 | scale |  | 4.11 | 16 | 7000 |  |
| manifesto-protectionism | 1000 | (1, 3) | 0.7 | 0.01 | rbf | 118.79 | scale |  | 19.86 | 23 | 3000 | yes |
| cap-sotu | 1000 | (1, 3) | 0.7 | 0.01 | linear | 2.57 | auto | balanced | 7.5 | 46 | 3000 | yes |
| manifesto-military | 500 | (1, 3) | 0.7 | 0.03 | linear | 1.84 | scale |  | 3.49 | 42 | 3000 | yes |
| coronanet | 500 | (1, 3) | 0.8 | 0.01 | linear | 6.79 | auto |  | 1.43 | 49 | 7000 |  |
| manifesto-8 | 500 | (1, 3) | 0.7 | 0.01 | linear | 2.57 | auto | balanced | 7.5 | 46 | 3000 | yes |
| cap-us-court | 500 | (1, 2) | 0.7 | 0.01 | linear | 15.47 | auto | balanced | 71.31 | 25 | 2000 |  |
| manifesto-morality | 500 | (1, 3) | 0.7 | 0.01 | linear | 2.57 | auto | balanced | 7.5 | 46 | 3000 | yes |
| sentiment-news-econ | 500 | (1, 3) | 0.7 | 0.03 | linear | 1.92 | scale |  | 1.11 | 41 | 2000 |  |
| manifesto-protectionism | 500 | (1, 3) | 0.8 | 0.03 | rbf | 3.31 | scale |  | 1.23 | 19 | 5000 | yes |
| cap-sotu | 500 | (1, 2) | 0.8 | 0.01 | poly | 520.17 | auto |  | 14.02 | 32 | 4000 | yes |
| manifesto-military | 100 | (1, 3) | 0.9 | 0.06 | linear | 7.15 | auto | balanced | 3.61 | 46 | 2000 | yes |
| coronanet | 100 | (1, 2) | 0.9 | 0.01 | sigmoid | 76.41 | scale | balanced | 11.67 | 45 | 6000 |  |
| manifesto-8 | 100 | (1, 3) | 0.7 | 0.03 | linear | 1.44 | scale |  | 1.02 | 41 | 1000 | yes |
| cap-us-court | 100 | (1, 3) | 0.9 | 0.03 | linear | 180.43 | scale |  | 40.17 | 19 | 2000 |  |
| manifesto-morality | 100 | (1, 2) | 0.8 | 0.06 | rbf | 2.85 | scale | balanced | 13.8 | 39 | 7000 | yes |
| sentiment-news-econ | 100 | (1, 2) | 0.7 | 0.03 | linear | 1.94 | scale |  | 19.29 | 12 | 2000 |  |
| manifesto-protectionism | 100 | (1, 2) | 0.8 | 0.01 | sigmoid | 9.17 | scale |  | 1.08 | 26 | 2000 | no |
| cap-sotu | 100 | (1, 2) | 0.8 | 0.03 | linear | 53.79 | scale | balanced | 1.12 | 33 | 2000 | yes |

Table 34 - Best hyperparameters logistic regression with TFIDF

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **dataset** | **sample** | **ngram** | **max\_df** | **min\_df** | **solver** | **C** | **class\_weight** | **max\_iter** | **warm\_start** | **context** |
| coronanet | 10000 | (1, 3) | 0.8 | 0.01 | liblinear | 5.04 |  | 327 | FALSE |  |
| manifesto-8 | 10000 | (1, 2) | 0.9 | 0.01 | sag | 1.04 |  | 99 | TRUE | yes |
| cap-sotu | 10000 | (1, 2) | 0.8 | 0.01 | saga | 8.14 |  | 579 | TRUE | yes |
| cap-us-court | 5426 | (1, 2) | 0.8 | 0.01 | saga | 772.47 | balanced | 825 | FALSE |  |
| coronanet | 5000 | (1, 2) | 0.8 | 0.01 | liblinear | 3.04 |  | 221 | FALSE |  |
| manifesto-8 | 5000 | (1, 2) | 0.9 | 0.01 | saga | 770.76 |  | 51 | TRUE | yes |
| cap-us-court | 5000 | (1, 2) | 0.8 | 0.01 | saga | 775.36 | balanced | 618 | TRUE |  |
| cap-sotu | 5000 | (1, 2) | 0.9 | 0.01 | saga | 12.96 |  | 583 | TRUE | yes |
| manifesto-military | 3970 | (1, 2) | 0.9 | 0.01 | lbfgs | 3.67 |  | 135 | FALSE | yes |
| manifesto-morality | 3188 | (1, 2) | 0.7 | 0.01 | sag | 2.35 |  | 871 | TRUE | yes |
| sentiment-news-econ | 3000 | (1, 2) | 0.8 | 0.01 | saga | 3.21 |  | 213 | FALSE |  |
| manifesto-military | 2500 | (1, 2) | 0.9 | 0.01 | sag | 3.97 |  | 884 | TRUE | yes |
| coronanet | 2500 | (1, 2) | 0.8 | 0.01 | saga | 3.69 |  | 294 | TRUE |  |
| manifesto-8 | 2500 | (1, 2) | 0.7 | 0.01 | sag | 3.37 |  | 568 | FALSE | yes |
| cap-us-court | 2500 | (1, 2) | 0.9 | 0.01 | lbfgs | 615.39 | balanced | 542 | TRUE |  |
| manifesto-morality | 2500 | (1, 2) | 0.9 | 0.01 | saga | 4.26 |  | 50 | FALSE | yes |
| sentiment-news-econ | 2500 | (1, 3) | 0.7 | 0.06 | saga | 5.76 |  | 252 | TRUE |  |
| cap-sotu | 2500 | (1, 2) | 0.7 | 0.01 | saga | 112.09 | balanced | 495 | FALSE | yes |
| manifesto-protectionism | 2116 | (1, 2) | 0.9 | 0.01 | lbfgs | 3.37 |  | 52 | FALSE | yes |
| manifesto-military | 1000 | (1, 2) | 0.8 | 0.01 | saga | 151.31 |  | 919 | TRUE | yes |
| coronanet | 1000 | (1, 3) | 0.9 | 0.01 | liblinear | 94.58 | balanced | 282 | TRUE |  |
| manifesto-8 | 1000 | (1, 2) | 0.9 | 0.01 | saga | 857.06 |  | 52 | FALSE | yes |
| cap-us-court | 1000 | (1, 2) | 0.7 | 0.01 | saga | 822.71 | balanced | 633 | FALSE |  |
| manifesto-morality | 1000 | (1, 2) | 0.9 | 0.01 | lbfgs | 57.92 |  | 52 | FALSE | yes |
| sentiment-news-econ | 1000 | (1, 2) | 0.9 | 0.01 | sag | 4.07 |  | 413 | TRUE |  |
| sentiment-news-econ | 1000 | (1, 3) | 0.7 | 0.03 | sag | 15.53 |  | 973 | FALSE |  |
| manifesto-protectionism | 1000 | (1, 2) | 0.9 | 0.01 | lbfgs | 896.33 |  | 903 | FALSE | yes |
| cap-sotu | 1000 | (1, 2) | 0.7 | 0.01 | saga | 112.09 | balanced | 495 | FALSE | yes |
| manifesto-military | 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-8 | 500 | (1, 2) | 0.7 | 0.01 | saga | 209.71 |  | 57 | FALSE | yes |
| cap-us-court | 500 | (1, 2) | 0.8 | 0.01 | saga | 772.47 | balanced | 825 | FALSE |  |
| manifesto-morality | 500 | (1, 2) | 0.9 | 0.01 | sag | 237.68 |  | 734 | FALSE | yes |
| sentiment-news-econ | 500 | (1, 3) | 0.9 | 0.06 | sag | 213.13 |  | 374 | TRUE |  |
| sentiment-news-econ | 500 | (1, 2) | 0.8 | 0.01 | saga | 72.12 |  | 72 | TRUE |  |
| manifesto-protectionism | 500 | (1, 2) | 0.7 | 0.01 | lbfgs | 741.03 |  | 391 | FALSE | yes |
| cap-sotu | 500 | (1, 2) | 0.9 | 0.01 | sag | 921.76 |  | 125 | TRUE | yes |
| manifesto-military | 100 | (1, 3) | 0.8 | 0.03 | newton-cg | 3.67 |  | 342 | TRUE | yes |
| coronanet | 100 | (1, 2) | 0.8 | 0.01 | saga | 775.36 | balanced | 618 | TRUE |  |
| manifesto-8 | 100 | (1, 2) | 0.9 | 0.01 | sag | 919.68 |  | 51 | TRUE | yes |
| cap-us-court | 100 | (1, 2) | 0.8 | 0.03 | liblinear | 950.46 | balanced | 639 | TRUE |  |
| manifesto-morality | 100 | (1, 2) | 0.8 | 0.01 | saga | 252.68 |  | 65 | TRUE | yes |
| sentiment-news-econ | 100 | (1, 2) | 0.8 | 0.01 | saga | 798.47 |  | 161 | TRUE |  |
| sentiment-news-econ | 100 | (1, 3) | 0.7 | 0.03 | lbfgs | 609.95 | balanced | 314 | TRUE |  |
| manifesto-protectionism | 100 | (1, 2) | 0.7 | 0.01 | sag | 390.7 |  | 50 | FALSE | yes |
| cap-sotu | 100 | (1, 3) | 0.8 | 0.03 | newton-cg | 279.31 | balanced | 164 | TRUE | yes |

# 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 35 - Training time comparison for a single model

|  |  |  |  |
| --- | --- | --- | --- |
| **algorithm** | **sample size** | **minutes training** | **hardware** |
| SVM | 100 | 0 | CPU (AMD Rome 7H12) |
| SVM | 500 | 0 | CPU (AMD Rome 7H12) |
| SVM | 1000 | 0 | CPU (AMD Rome 7H12) |
| SVM | 2500 | 0 | CPU (AMD Rome 7H12) |
| SVM | 5000 | 0,5 | CPU (AMD Rome 7H12) |
| SVM | 10000 | 1 | CPU (AMD Rome 7H12) |
| logistic regression | 100 | 0 | CPU (AMD Rome 7H12) |
| logistic regression | 500 | 0 | CPU (AMD Rome 7H12) |
| logistic regression | 1000 | 0 | CPU (AMD Rome 7H12) |
| logistic regression | 2500 | 0 | CPU (AMD Rome 7H12) |
| logistic regression | 5000 | 0,25 | CPU (AMD Rome 7H12) |
| logistic regression | 10000 | 0 | CPU (AMD Rome 7H12) |
| BERT-base-nli | 100 | 3,25 | GPU (A100) |
| BERT-base-nli | 500 | 4,75 | GPU (A100) |
| BERT-base-nli | 1000 | 7 | GPU (A100) |
| BERT-base-nli | 2500 | 11 | GPU (A100) |
| BERT-base-nli | 5000 | 23,5 | GPU (A100) |
| BERT-base-nli | 10000 | 45,33 | GPU (A100) |
| BERT-base | 100 | 1,62 | GPU (A100) |
| BERT-base | 500 | 6,25 | GPU (A100) |
| BERT-base | 1000 | 12,38 | GPU (A100) |
| BERT-base | 2500 | 23,71 | GPU (A100) |
| BERT-base | 5000 | 41,5 | GPU (A100) |
| BERT-base | 10000 | 69 | GPU (A100) |

# Bibliography

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.

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.

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.

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.

1. https://github.com/MoritzLaurer/less-annotating-with-bert-nli [↑](#footnote-ref-1)
2. https://manifesto-project.wzb.eu/ [↑](#footnote-ref-2)
3. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/MXKRDE [↑](#footnote-ref-3)
4. https://www.comparativeagendas.net/datasets\_codebooks [↑](#footnote-ref-4)
5. https://www.comparativeagendas.net/pages/master-codebook [↑](#footnote-ref-5)
6. https://www.comparativeagendas.net/datasets\_codebooks [↑](#footnote-ref-6)
7. https://www.coronanet-project.org/ [↑](#footnote-ref-7)
8. The full script written in Python is available on our GitHub repository: https://github.com/MoritzLaurer/less-annotating-with-bert-nli [↑](#footnote-ref-8)
9. https://optuna.readthedocs.io/en/stable/ [↑](#footnote-ref-9)
10. https://optuna.readthedocs.io/en/stable/ [↑](#footnote-ref-10)
11. 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> [↑](#footnote-ref-11)
12. 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. [↑](#footnote-ref-12)
13. Note that we tested the warmup ratio hyperparameter only for the sample sizes above 2500 for BERT-NLI. It would be interesting to also test it for other sample sizes, but we avoided repeating the hyperparameter searches to save computational costs. [↑](#footnote-ref-13)
14. https://universaldependencies.org/u/pos/ [↑](#footnote-ref-14)