Diversity Over Size: On the Effect of Sample and Topic Sizes for Topic-Dependent Argument Mining Datasets

Benjamin Schiller[a,b], **Johannes Daxenberger**[a], **Andreas Waldis**[b,c] and **Iryna Gurevych**[b]
[a]summetix GmbH, [b]Ubiquitous Knowledge Processing Lab,
Department of Computer Science, Technical University of Darmstadt, and
[c]Information Systems Research Lab, Lucerne University of Applied Sciences and Arts
[a]{schiller, daxenberger}@summetix.com,
[b]www.ukp.tu-darmstadt.de,
[c]www.hslu.ch

Abstract

Topic-Dependent Argument Mining (TDAM), that is extracting and classifying argument components for a specific topic from large document sources, is an inherently difficult task for machine learning models and humans alike, as large TDAM datasets are rare and recognition of argument components requires expert knowledge. The task becomes even more difficult if it also involves stance detection of retrieved arguments. In this work, we investigate the effect of TDAM dataset composition in few- and zero-shot settings. Our findings show that, while fine-tuning is mandatory to achieve acceptable model performance, using carefully composed training samples and reducing the training sample size by up to almost 90% can still yield 95% of the maximum performance. This gain is consistent across three TDAM tasks on three different datasets. We also publish a new dataset[1] and code[2] for future benchmarking.

1 Introduction

Topic-Dependent Argument Mining (TDAM) is the task of extracting argument components in documents or document collections (Lauscher et al., 2022). Topic-dependence (or, as Stab et al. (2018) refer to it, *information-seeking* argument mining) means that argument components are directed towards a given topic (Ein-Dor et al., 2020; Shnarch et al., 2018). The topic is used in two ways: by a machine learning model to learn topic-relevance and as a query to retrieve input documents for automatic argument search (Daxenberger et al., 2020).

While LLMs (large language models) show astounding results (Touvron et al., 2023; OpenAI, 2024), task-related datasets are still important to improve model performance (Dettmers et al., 2023; Lv et al., 2024; Liu et al., 2022; van der Meer et al., 2022) and decrease certain undesirable behaviours (Ouyang et al., 2022; Askell et al., 2021) via fine-tuning, and to provide curated data for evaluation purposes. To assemble large amounts of training samples, it is common to use non-experts to annotate datasets. However, in contrast to a task like sentiment analysis, the task of identifying arguments is not naturally understood by non-experts and, due to pitfalls like commonly used fallacies (Habernal et al., 2018), needs a thorough training phase and strict quality control of the crowdsourcing process. Hence, crowdsourcing datasets for TDAM is not only time-consuming but also expensive, as it requires a large number of workers per sample for satisfactory agreement. For instance, Stab et al. (2018) report a sum of $2,774 for the annotation of 25,492 samples, requiring seven annotators to reach a satisfying inter-annotator agreement.

Due to the efficacy of transformers (Vaswani et al., 2017), datasets for TDAM (as for many other tasks) have grown in size over recent years (Stab et al., 2018; Shnarch et al., 2018; Rinott et al., 2015; Aharoni et al., 2014). Recent datasets for TDAM contain up to 30,000 samples (Ein-Dor et al., 2020). However, relying on large datasets has several disadvantages: (1) it is impractical to label such large datasets by experts, (2) crowdsourcing them is costly, and (3) training (as well as tuning) takes longer and adds to the cost.

To tackle those disadvantages, we study if and how dataset sizes for TDAM can be reduced and what the *composition* (total number of topics, samples, and samples per topic) of these datasets should be to train high-performing models. In contrast to simpler text classification tasks with a single input (e.g. document categorization or sentiment analysis), creating datasets for cross-topic TDAM is more complicated, as it requires controlling two or more inputs (e.g. topic and argument component) and a diverse choice of topics, which we show in this work.

[1]https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/4353
[2]https://github.com/UKPLab/argument-topic-diversity

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 10870–10887
November 12-16, 2024

Our work is motivated by few-shot learning (Wei et al., 2022; Schick and Schütze, 2021; Rücklé et al., 2020; Vinyals et al., 2016) and diversity sampling (Larson et al., 2019; Katharopoulos and Fleuret, 2018; Chang et al., 2017) approaches. Larson et al. (2019) show that *unique* samples (similar to *outliers* in Chang et al. (2017)), i.e. samples that differ strongly in structure or content from other samples, can increase model robustness. Thus, in addition to relying on models that are able to learn with fewer samples, we increase the diversity of samples in our dataset by integrating a large number of distinct topics (i.e., outliers) and, in turn, aim to increase the robustness of our models.

As a testbed, we create a benchmark with two datasets that have an equal number of training samples and only differ in the number of topics and samples per topic. We research the influence of these two compositional parameters on model performance and costs of the annotation process, showing that a largely increased number of topics improves model performance by up to 4.1pp in this scenario. We verify findings from the benchmark datasets on two TDAM datasets from different domains with a slightly different task and find that we can save up to almost 90% of the annotation costs if we are willing to sacrifice 5% of the maximum model performance.

Our contributions are as follows: (1) We create a new dataset for TDAM which differs from an existing TDAM dataset, namely the UKP Sentential Argument Mining Corpus (UKP Corpus) (Stab et al., 2018), only in the number of topics and samples per topic, allowing for a deeper analysis of this task and assumptions on how future datasets can be composed (diversity sampling), (2) we analyze zero- and few-shot experiments on the new dataset, giving recommendations on efficient dataset composition, (3) we evaluate findings on dataset efficiency on two different TDAM tasks from another domain, and (4) we present state-of-the-art results on the UKP Corpus.

2 Related Work

TDAM The task of *Discourse-level* Argument Mining aims to classify argument components (Rocha et al., 2023; Ajjour et al., 2017; Stab and Gurevych, 2014; Goudas et al., 2014) and their relations (Eger et al., 2017; Nguyen and Litman, 2016) within isolated documents. *Topic-Dependent* Argument Mining (Lauscher et al., 2022), however, describes the task of searching large, heterogeneous document collections for argument components relevant to a given topic (Ein-Dor et al., 2020; Stab et al., 2018; Shnarch et al., 2018). In this work, we will focus on the latter instead of extracting components like claims and premises or their relations from single documents.

Dataset composition The growth of sample sizes in TDAM datasets seems to be a necessity to cover wider ranges of topics and, thus, to support better cross-topic and cross-domain performance (Ein-Dor et al., 2020; Stab et al., 2018; Shnarch et al., 2018; Rinott et al., 2015; Aharoni et al., 2014). As this, in consequence, increases the annotation and training costs of the models, we aim to uncover low-effort methods for TDAM datasets that help to keep the number of training samples as low as possible while reaching similar performance. Ajjour et al. (2023) discover that many TDAM datasets, even those with large amounts of samples, do mostly cover topics that frequently appear in forums, but leave out many less-frequently discussed areas. We argue that using too many samples per topic in TDAM datasets is a waste of financial resources and focusing on only a few, frequently discussed topics limits the capability of models to generalize well in cross-topic experiments. In this work, we propose a different compositional structure for future TDAM datasets.

Model learning techniques The intuition of keeping the number of training samples—and hence, costs and annotation effort—low, has attracted research that focuses on techniques enabling models to learn with less data. With regard to models, one of the most impactful designs in recent years are transformer (Vaswani et al., 2017) that are pre-trained on large amounts of text with unsupervised learning techniques (Liu et al., 2019b; Devlin et al., 2019). These LLMs show remarkable results on few-shot learning (Gao et al., 2021; Schick and Schütze, 2021) and zero-shot learning (Wei et al., 2022; Rücklé et al., 2020; Radford et al., 2019) tasks. One form of zero-shot learning that gained a lot of attention due to its astounding performance is *prompting* (in-context learning), where a pre-trained model is not fine-tuned and, in addition to the actual input, is only given exemplary inputs (for instance, prepended to the actual input) at inference (Brown et al., 2020). Two other and

older techniques used to reduce training sample sizes are transfer and multi-task learning, which have also been successfully combined with LLMs (Schiller et al., 2021; Liu et al., 2019a).

Benchmarks and diversity sampling In contrast to most of these techniques that concentrate on adapting model architectures in a way such that models are able to learn with few or no samples, we focus on benchmarking efficient compositional structures of TDAM datasets. In recent years, other benchmarks consisting of multiple datasets have been published (Schiller et al., 2021; Wang et al., 2019a,b) with the aim to standardize performance reports for machine learning models and, hence, allowing new model architectures to compete against each other and to make the results comparable. Arakelyan et al. (2023) show improvements on a benchmark for the task of stance detection by sampling a subset with a deep, unsupervised topic model and training on the subset with a contrastive objective. We, however, aim to adapt and benchmark the composition of TDAM datasets we train the models with, such that existing models (without any modification) can exploit it. Our decision on how to ensemble the datasets we use in this work draws on insights of diversity sampling research which shows that models can profit from datasets with high diversity, i.e. containing samples that differ strongly from each other (Larson et al., 2019; Katharopoulos and Fleuret, 2018; Chang et al., 2017). While other work in the area of TDAM has scratched on the topic of diversity by increasing the number of used topics (Ein-Dor et al., 2020), there has been no work that we know of, dedicated on determining the ideal dataset composition for TDAM datasets.

3 Data

For our dataset composition benchmark, we first need two datasets that only differ in the aforementioned dimensions of *number of topics* and *number of samples per topic* but are otherwise similarly composed. We use one existing dataset (see Section 3.1) and base a new dataset (see Section 3.2) on it with a composition better fit for diversity sampling and few-shot learning. We evaluate our hypotheses on dataset composition for TDAM on two more TDAM datasets (see Sections 3.3 and 3.4) with slightly different learning tasks. Statistical information about all datasets as well as examples can be found in Tables 1 and 2. Information about dataset licenses are listed in Appendix B.

3.1 UKP Corpus

As opposed to other TDAM datasets (Ein-Dor et al., 2020; Shnarch et al., 2018; Rinott et al., 2015; Aharoni et al., 2014), the UKP Corpus has two main advantages: First, it includes stance labels, which are an important additional information to categorize mined arguments and can be further processed for tasks like fake news detection (Hanselowski et al., 2018). Second, the dataset is from heterogeneous data sources and models real-world scenarios better than taking only samples from a single source. It consists of eight topics with a total of 25,492 samples, which are pairs of a short topic and a single sentence, labeled with *argument for* (pro), *argument against* (con), or *no argument* (none). As described by Stab et al. (2018), a sentence is only labeled as pro or con argument, if it holds evidence for why the sentence supports or opposes the topic. If the sentence holds no such evidence or is unrelated to the topic, it is labeled as *no argument*. We split the dataset by taking all samples of five topics for training, of one topic for development, and of two topics for testing. To allow for a fair comparison, we downsample the number of samples in the training set (equally for each topic) to fit the total number of training samples generated for our newly created Few-Shot-150T Corpus (FS150T-Corpus).

3.2 FS150T-Corpus

Due to its aforementioned advantages, we decide to base our new dataset on the UKP Corpus. We follow the exact guidelines and data crawling strategy used for the UKP Corpus and crowdsource 21,600 samples over 150 controversial topics with exactly 144 samples for each topic (see Appendix A.1). The composition of our dataset is therefore ideal for few-shot learning, as we have the same number of samples for each topic to easily scale up and down from 0 to 144. Moreover, we have a large amount of topics to scale diversity up and down. The topics are a collection of controversial subjects from multiple domains like politics, technology, economy, and do not intersect with topics from the UKP Corpus (see Appendix Table 8). We randomly pick 10 of the topics for our development set (1,440 samples) and 20 topics for our test set (2,880) samples, leaving 120 topics for the training set (17,280 samples).

Datasets	# Topics	# Samples				Classes
		Train	Dev	Test	Total	
FS150T-Corpus (ours)	150	17,280	1,440	2,880	21,600	pro (19%), con (19%), none (62%)
UKP Corpus (Stab et al., 2018)	8	17,280	2,475	1,249	5,481	pro (19%), con (24%), none (56%)
IAM-Corpus (Cheng et al., 2022)	100	9,678	7,057	7,065	23,800	support (11%), contest (10%), no relation (79%)
IBM-Corpus (Ein-Dor et al., 2020)	221	22,396	2,954	4,079	29,429	evidence (23%), no evidence (77%)

Table 1: Splits, classes, and class distributions for all used datasets.

Dataset	Domain	Topic	Sentence	Class
FS150T-Corpus	Web Search	electronic cigarettes	Currently, there is no scientific evidence confirming that electronic cigarettes help smokers quit smoking cigarettes.	contra
		renewable energy	Installation is quick and homeowners can be enjoying solar energy in a matter of days.	pro
UKP Corpus	Web Search	nuclear energy	It is pretty expensive to mine, refine and transport uranium.	contra
		gun control	Gun control laws would reduce the societal costs associated with gun violence.	pro
IAM-Corpus	Encyclopedia	Should you restrict reality TV	They involve extreme competition which drains children; it takes away their innocence.	contest
		Should boxing be banned	With a careful and thoughtful approach, boxing quite can be beneficial to health.	support
IBM-Corpus	Encyclopedia	We should ban organic food	Like local food systems, organic food systems have been criticized for being elitist and inaccessible.	argument

Table 2: All datasets used in this work with the general domain the data origins from and data samples with topic, sentence, and annotated labels (class).

3.3 IAM-Corpus (IAM-Corpus)

The IAM-Corpus is built upon the data from "Task 1: Claim Extraction" by Cheng et al. (2022). The original data is based on 123 debating topics and 1,010 related articles from English Wikipedia. One sample consists of a topic and a sentence from an article. Each pair has one of three possible labels attached: *support*, *contest*, or no *relation*. Due to the massive imbalance of the none-arguments in the original training split (93%), we have to downsample them in a way that the model is able to pick up the other two classes. We randomly pick samples until we reach a class distribution of 22%/18%/60% (support/contest/no relation) in the training set. We leave the dev and test sets untouched from the original, which also makes this dataset the only in-topic dataset (as opposed to cross-topic datasets which have no overlapping topics between the dataset splits). The modified training data set contains all 100 topics, the original dev and test sets contain 62 topics and 63 topics.

3.4 IBM-Corpus (IBM-Corpus)

The IBM-Corpus is based on the publicly available dataset constructed by the authors in Ein-Dor et al. (2020). The dataset consists of almost 30,000 *motion*-sentence samples and each sample has a score between 0 and 1 that either denotes a sentence to be rather an evidence for the related motion or not. Motions are described as high-level claims, e.g. "Capitalism brings more harm than good". The sentences are extracted from English Wikipedia. Following the authors' experimental setup, we set a threshold at 0.6 for the score to define two class labels *evidence* and *no evidence*. We take 35 random topics to form the test set, 20 random topics to form the dev set, and 166 topics to form the training set. In contrast to all other datasets, this one has only two class labels and the largest number of topics (here: motions) and samples.

4 Method

To investigate the optimal composition of TDAM datasets, we conduct sample, topic, and dataset experiments, which we elaborate in the following.

4.1 Sample experiments

We investigate on how many training samples per topic are necessary to reach *acceptable* and maximum performance. We start our experiments with 0 training samples per topic, i.e. untrained model performance (zero-shot) and increase the number of samples in small steps, ending with all samples available for each topic. We define acceptable performance by reaching at least 95% of the highest performance on a test set, measured over all sample

experiments for a given model. We define maximum performance for a model by the highest value for the given metric on a test set, regardless of the number of training samples used to reach it.

4.2 Topic experiments

We analyze how many topics are needed to generalize well in cross-topic experiments by choosing a set of training sample sizes (960; 1,440; 2,880; 5,760) and fixing them while increasing the topics. The topics are increased in steps of 5 and end with the maximum number of topics available for a training set. Since fixing the number of topics and sample sizes requires a certain amount of samples available for each topic in the training set, experiments with larger sample sizes may start at larger topic sizes. For instance, fixing 10 topics and 960 samples only requires 96 samples per topic in the training set, while fixing 10 topics and 5,760 samples already requires 576 samples per topic. The more topics we include for a fixed sample size, the less samples per topic are available. We aim to find out whether or not using many topics (diversity sampling) is beneficial for cross-topic performance.

4.3 Dataset experiments

We investigate if we can reach higher performance by training on a dataset with few topics but many samples per topic (UKP Corpus) or on a dataset with many topics but few samples per topic (FS150T-Corpus). For the benchmark experiments, we leverage both supervised models and re-train them in the following to setups:

- Training and tuning on the UKP Corpus and show results on the test sets of both corpora.
- Training and tuning on the FS150T-Corpus and show results on the test sets of both corpora.

If either of the two variants performs better in both experiments, we know the dataset composition that should be preferred for the task of TDAM. In any other case, our assumption that few samples combined with many topics is the superior dataset composition (i.e. better cross-topic performance) on TDAM datasets is refuted.

5 Models

We use four models: ERNIE 2.0 as a strong and fast to train language model, FLAN-T5 XL as an LLM option that was trained on massive amounts of data, and two state-of-the-art chat models for zero-shot in-context learning experiments. Details on fine-tuning parameters are described in Appendix A.2.

5.1 ERNIE 2.0

As medium-sized model (110M parameters), we use ERNIE 2.0 (Sun et al., 2020) which was pre-trained in a continual multi-task learning fashion on several word-, structure-, and semantic-aware tasks (but not on TDAM tasks) and showed state-of-the-art performance when fine-tuned on tasks of the GLUE Benchmark (Wang et al., 2019b). The data for the pre-training tasks was automatically generated with text extracted from encyclopedias, books, dialog, and discourse relation datasets. As these tasks have similar properties to TDAM, we expect to benefit from the pre-training through a higher maximum performance. Moreover, we anticipate that the specific pre-training enables the model to bootstrap performance on few-shot learning.

5.2 FLAN-T5 XL

We use FLAN-T5 XL (Chung et al., 2024) as a *large* language model for our experiments. It is a variant of the T5 model (Raffel et al., 2020) which was fine-tuned on 1.8K instruction tasks. As this model has an encoder-decoder structure, we remove the decoder and use a classification head for our tasks, which leaves around 1.3B parameters. Fine-tuning is done with LoRa (Hu et al., 2022) to reduce training time.

5.3 LLama2-70B, ChatGPT

As strong, zero-shot baselines, we use Llama2-70B (Touvron et al., 2023) and ChatGPT (OpenAI, 2023) in our experiments. The prompts used for each dataset and the specific model versions can be found in Appendix A.2.2.

6 Topic & Sample Experiments and Evaluation

We run all experiments over six seeds and report the average F_1 macro on the test set for the three seeds with the highest F_1 macro measured on the development set (see Appendix A.2).[3]

6.1 Sample Experiments

Figures 1-3 (see also Appendix C) show the performance gains of all tested models with increasing sample sizes per topic (samples uniformly dis-

[3]With low sample sizes, the models sometimes fail on some of the seeds and distort the results.

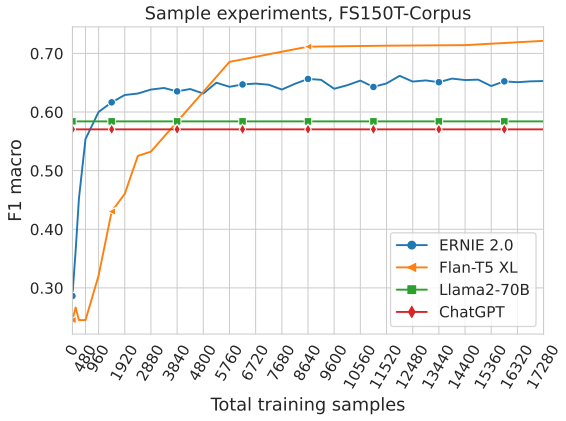

Figure 1: Sample experiments on the FS150T-Corpus

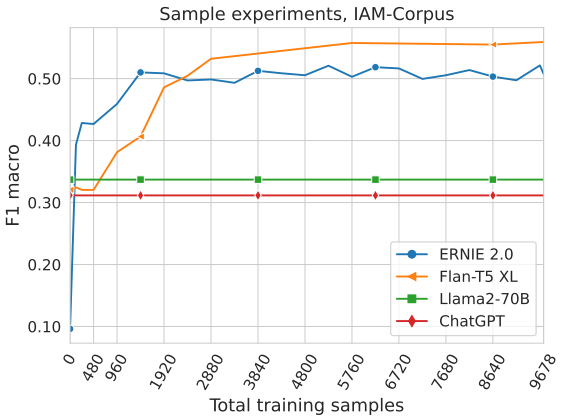

Figure 2: Sample experiments on the IAM-Corpus

tributed over all training topics)[4]. As two strong zero-shot baselines, we show results with Llama2-70B and ChatGPT.

For all datasets, we observe that FLAN-T5 XL struggles with small sample sizes, whereas ERNIE 2.0 shows good performance early on. ERNIE 2.0 reaches >0.60 F_1 macro on the FS150T-Corpus at 960 samples (FLAN-T5 XL: >2,800 samples), >0.70 F_1 macro on the IBM-Corpus at 960 samples (FLAN-T5 XL: 2,400 samples), and >0.50 F_1 macro on the IAM-Corpus at 1,440 samples (FLAN-T5 XL: 2,400 samples). However, when looking at maximum performance, FLAN-T5 XL eventually outperforms ERNIE 2.0 on all datasets on up to 6pp (percentage points) in F_1 macro. Both models require all training data to reach their maximum performance, except ERNIE 2.0 on the FS150T-Corpus where it peaks at 69% of the training data and FLAN-T5 XL on the IBM-Corpus where it still requires 90% of the training data. Llama2-70B and ChatGPT outperform the other models on all datasets on zero-shot experiments. However, they both loose their advantage after 500-2,000 training samples eventually, depending on the dataset. Interestingly, on our FS150T-Corpus that is especially designed for few-shot learning, both zero-shot baselines are the most competitive to our supervised learning models.

We also investigate how much data we really need to reach *acceptable* performance, i.e., 95% of the maximum performance of a model (see Table 3). For all three datasets, ERNIE 2.0 only needs a maximum of 15% of the data to reach this performance. For the FS150T-Corpus, it only needs 11%

[4]Increased step size of 2,880 for FLAN-T5 XL from 2,880 samples onwards.

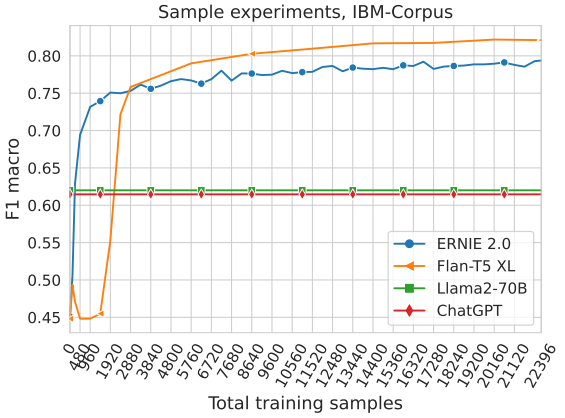

Figure 3: Sample experiments on the IBM-Corpus

Dataset	Model	# Samples: max performance	F_1 macro	# Samples: 95% of max performance
FS150T-Corpus	FLAN-T5 XL	17,280 (100%)	.7214 ± .0064	5,760 (33%)
	ERNIE 2.0	**12,000** (69%)	.6617 ± .0048	**1,920** (11%)
	ChatGPT	0	.5700	-
	Llama2-70B	0	.5608	-
	Majority	0	.2451	-
IAM-Corpus	FLAN-T5 XL	9,678 (100%)	**.5591** ± .0234	2,880 (30%)
	ERNIE 2.0	**6,720** (99%)	.5213 ± .0070	**1,440** (15%)
	ChatGPT	0	.2890	-
	Llama2-70B	0	.2920	-
	Majority	0	.3204	-
IBM-Corpus	FLAN-T5 XL	**20160** (90%)	**.8210** ± .0043	5760 (26%)
	ERNIE 2.0	22,396 (100%)	.7937 ± .0015	**3,360** (15%)
	ChatGPT	0	.6150	-
	Llama2-70B	0	.6210	-
	Majority	0	.4481	-

Table 3: Sample experiment results with training samples required for highest performance, highest performance in F_1 macro, and number of training samples required to reach 95% of the highest performance.

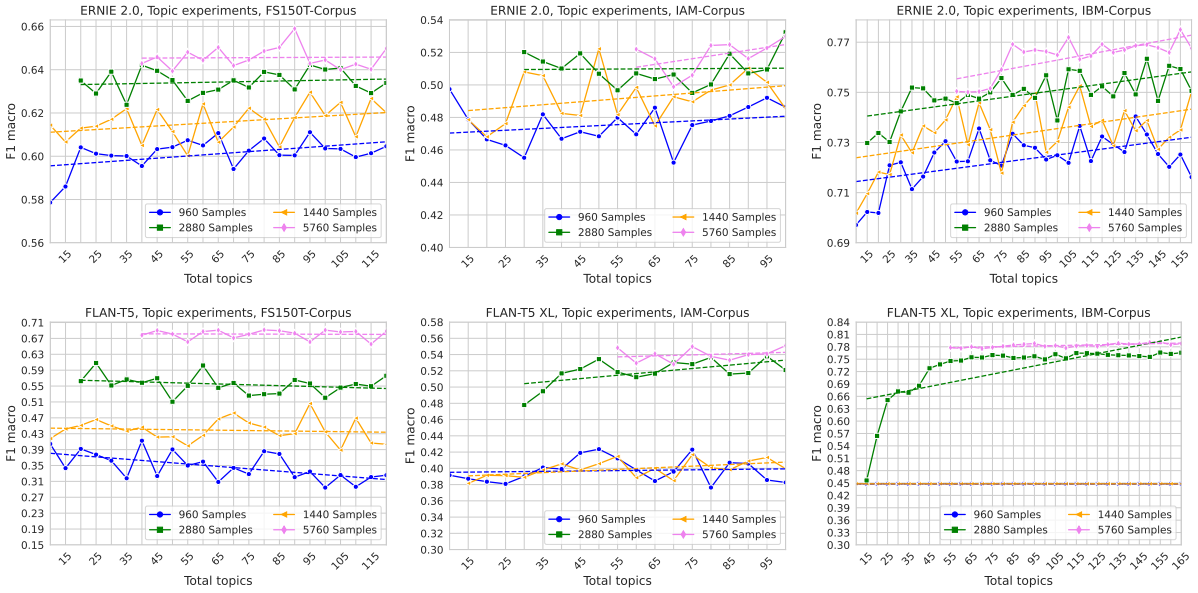

Figure 4: Topic experiments for FS150T-/IAM- and IBM-Corpus on ERNIE 2.0 and FLAN-T5 XL and in F_1 macro.

(with as few as 16 samples per topic). Hence, when using ERNIE 2.0, we can almost drop 90% of the data (if the dataset has a composition as proposed in this work). For FLAN-T5 XL, which has difficulties with smaller sample sizes on the datasets, it still only needs 26-33% to reach 95% of its maximum performance on the datasets, which could also reduce the necessary data size by almost 70%.

6.2 Topic Experiments

For the experiments, we fine-tune ERNIE 2.0 and FLAN-T5 XL on four different training sample sizes from 960 to 5,760. Topics and samples for each run are picked randomly.

We have a deeper look into how model performance changes for all datasets if the number of topics is increased while the training sample size is fixed (see Figure 4 and Appendix C). By increasing the number of topics, we also increase the diversity of the training set through adding outliers, instead of just picking more samples with similar content from the existing topics.

Similar to the increased robustness observed by Larson et al. (2019), ERNIE 2.0 shows an upward trend (dashed lines) for all datasets and sample sizes. For the FS150T-Corpus, we observe an upward trend with up to to 1.1pp on 960 samples, decreasing when more training samples are used. The largest upward trend for IAM-Corpus is reached on 5.760 samples (2.2pp) and for the IBM-Corpus, the largest upward trend is reached on 1,440 samples (1.9pp). Hence, ERNIE 2.0 is able to leverage diverse topic distribution in all tested scenarios. For the much larger FLAN-T5 XL, the impact of topic diversity is mixed. For experiments with 1,440 and less samples, we observe either no clear trend or even a negative trend (also slightly for 2,880 samples on the FS150T-Corpus). Using more samples, we observe a more positive trend. We assume these mixed results are due to two reasons: First, FLAN-T5 XL has seen much larger quantities of data initially in its pre-training as compared to ERNIE 2.0, which makes it harder to add even more diverse training data. Second, the large model size leads to unstable results with low sample sizes—we observe this for the IBM-Corpus in the topic experiments and, generally, in the sample experiments (see Figures 1-3). Hence, we conclude that the significance of the experiments with FLAN-T5 XL on smaller sample sizes is low.

7 Cross-Dataset Experiments and Evaluation on Benchmark Dataset

We tune two models based on ERNIE 2.0 and FLAN-T5 XL: one model trained and tuned on the FS150T-Corpus and one model trained and tuned on the UKP Corpus. We show evaluation results for both models on both corpora in Table 4.

Our baseline setting is when training, tuning, and testing happens on the same dataset. For that setup, FLAN-T5 XL shows .7532 and .7343 F_1

		Test on UKP Corpus			Test on FS150T-Corpus
	Test topics:	MW	SU	all	all
TACAM-BERT Base*	Train & tune on UKP Corpus	.4900	.6900	-	-
TACAM-BERT Large*	Train & tune on UKP Corpus	.6900	.6900	-	-
ERNIE 2.0	Train & tune on UKP Corpus	.6777 ± .0118	.7149 ± .0093	.6980 ± .0085	.6292 ± .0072
	Train & tune on FS150T-Corpus	**.7058** ± .0046	**.7406** ± .0029	**.7243** ± .0048	**.6585** ± .0036
FLAN-T5 XL	Train & tune on UKP Corpus	.7333 ± .0095	.7881 ± .0114	.7532 ± .0132	.6917 ± .0132
	Train & tune on FS150T-Corpus	**.7574** ± .0055	**.8270** ± .0023	**.7944** ± .0021	**.7343** ± .0037

Table 4: Dataset experiment results with ERNIE 2.0 and FLAN-T5 XL, comparing results on the FS150T-Corpus and UKP Corpus. As a baseline for the UKP Corpus, we use TACAM-BERT (*work by Fromm et al. (2019)). MW=Minimum Wage, SU=School uniforms.

macro for the UKP Corpus and the FS150T-Corpus, respectively. Training and tuning FLAN-T5 XL on the UKP Corpus and then evaluating it on the FS150T-Corpus shows worse performance (.6917 F_1 macro) than training and tuning a model on the actual FS150T-Corpus, which is the expected outcome. However, training and tuning FLAN-T5 XL on the FS150T-Corpus shows the best results on the UKP Corpus test set (.7944 F_1 macro). The same observation is made with ERNIE 2.0, just with lower overall F_1 macro scores. Hence, using the FS150T-Corpus for training performs best on the test sets of both corpora. We assume that training on just a few topics and fitting the model with a large number of training samples to those topics will not prepare it enough to generalize well — not even for a massively pre-trained model like FLAN-T5 XL. Training on many diverse topics, however, will add generalizability to the model and reduce the risk to over-fit to a small range of specific topics, that is, it will also learn that topics can come from a much larger variety within the embedding space.

We also compare the models' performances to the current state-of-the-art (for which topic-wise results are available) on the UKP Corpus. TACAM-BERT Base (Fromm et al., 2019), with a number of parameters comparable to ERNIE 2.0, performs 21.6pp and 5.1pp lower in F_1 macro for the test topics *minimum wage* and *school uniforms*. The much larger TACAM-BERT Large (three times the number of parameters) still underperforms ERNIE 2.0 by 1.6pp and 5.1pp.

8 Conclusion

We create a new dataset that enables to benchmark the composition of TDAM datasets. Experiments show that having many topics in combination with few samples per topic can improve model performance by 4.1pp in cross-dataset experiments and also reaches a new state-of-the-art on the UKP Corpus (see Section 7).

Recommendations for dataset composition: Overall, we observe a positive trend in performance when the number of training topics is increased and a medium-sized LM is used (ERNIE 2.0), but mixed results with an LLM (FLAN-T5 XL), which we attribute to the extensive and diverse pre-training and generally more unstable results on smaller sample sizes (see Section 6.2). While the topic experiments do not show a drastic increase of accuracy, it can be an easy way to improve the performance and usually comes without additional costs. Hence, if there is a sample limit for a planned dataset, we can increase a medium-sized model's performance by composing the dataset with more topics. As we tested up to 160 topics on all datasets, we assume this to be a good choice for training data sizes ranging from ~1,000-6,000 samples but can become less relevant if a large model like FLAN-T5 XL is used. However, in many scenarios where inference speed and operating costs are decisive, a smaller model like ERNIE 2.0 with carefully sampled training data might be the preferred choice.[5]

When choosing our proposed dataset composition (FS150T-Corpus) in combination with ERNIE 2.0 (pre-trained on several word-, structure-, and semantic-aware tasks), we can reduce the training sample size by almost 90% (to 1,920 samples), still reach 95% of the maximum performance and, in turn, decrease the annotation costs of the train set by $2,323 to only $290 (see Appendix A.1) for a dataset created in the composition proposed in this work (see Section 6.1). Although showing difficulties on small samples sizes, we can still reduce the sample size by 67% with FLAN-T5 XL. We observe the same trend on the other two datasets: on the IBM-Corpus, 15% of the training data with ERNIE 2.0 and 26% with FLAN-T5 XL are sufficient to reach 95% of the maximum performance; on the IAM-Corpus, 15% with ERNIE 2.0 and

[5]In comparison on an NVIDIA A10, FLAN-T5 XL can predict 34 samples per second with a batch size of 1 and takes 5,638MB of GPU memory, whereas ERNIE 2.0 can predict 111 samples per second and only takes 860MB.

30% with FLAN-T5 XL are sufficient. This clearly challenges the trend to develop larger datasets for TDAM. Following our proposed dataset composition makes low-budget productions of high-quality TDAM datasets possible, contributing to a more diverse landscape of those datasets.

Generalization to other tasks: While we test our approach on TDAM only, work on related tasks also shows performance improvements when focusing on dataset diversity. For instance, Arakelyan et al. (2023) show that higher performance can be reached if subsets of stance detection datasets are sampled for diversity with an unsupervised topic model and used for training. Sultan et al. (2020) use a transformer-based question generator and conclude that more diverse questions lead to a higher performance on downstream Question-Answering tasks. Yadav et al. (2024) also show improvements for Question-Answering by generating question-answer pairs with a focus on diversity conditions like spacial aspects, question types and entities. Similar to our findings, they also show that performance gains are highest in low-resource scenarios. Hence, we speculate that a well-architected dataset composition, based on diversity, can also lead to a high performance on Question-Answering and pure Stance Detection datasets.

We publish our newly created dataset[1] and code[2], allowing for further benchmarking experiments to develop the design of future TDAM datasets.

Limitations

Our experiments focus on datasets for TDAM only. While we would expect other tasks with datasets of similar composition (for instance, Stance Detection or Question-Answering, as discussed in Section 8) to also profit from our findings, we have not tested this and can only make claims based on our experiments for TDAM.

Our sample experiments only cover sample sizes from 0 to 22,396 training samples and a step size of 480-2,880 samples, depending on the model. Hence, we can not rule out the possibility of higher F_1 macro or other derivations from the observable trend with more than 22,396 training samples, nor can we rule out the possibility that we have missed a certain dip or peak due to our chosen step size. Similarly for our topic experiments—while there is a trend to higher performance with more topics, it is unclear how this trend develops with more than 160 topics (for instance, if the model shows a saturation with regard to topics or if more topics would even have a negative impact).

Lastly, when comparing the maximum performance of our medium sized model choice with our LLM choice, there is an obvious gap in performance to observe. There is, however, a clear downside of using LLMs when it comes to operation cost and speed (see Section 8), which can be a crucial factor in many scenarios.

Acknowledgements

This work has been funded by the LOEWE Distinguished Chair "Ubiquitous Knowledge Processing", LOEWE initiative, Hesse, Germany (Grant Number: LOEWE/4a//519/05/00.002(0002)/81) and by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 01UP2229B (KoPoCoV).

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland. Association for Computational Linguistics.

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128, Copenhagen, Denmark. Association for Computational Linguistics.

Yamen Ajjour, Johannes Kiesel, Benno Stein, and Martin Potthast. 2023. Topic ontologies for arguments. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1411–1427, Dubrovnik, Croatia. Association for Computational Linguistics.

Erik Arakelyan, Arnav Arora, and Isabelle Augenstein. 2023. Topic-guided sampling for data-efficient multi-domain stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13448–13464, Toronto, Canada. Association for Computational Linguistics.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *ArXiv preprint*, abs/2112.00861, Version 3.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Haw-Shiuan Chang, Erik G. Learned-Miller, and Andrew McCallum. 2017. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *Advances in Neural Information Processing Systems*, volume 30, pages 1002–1012. Curran Associates, Inc.

Liying Cheng, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang, and Luo Si. 2022. IAM: A comprehensive and large-scale dataset for integrated argument mining tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2277–2287, Dublin, Ireland. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Johannes Daxenberger, Benjamin Schiller, Chris Stahlhut, Erik Kaiser, and Iryna Gurevych. 2020. Argumentext: Argument classification and clustering in a generalized search scenario. *Datenbank-Spektrum*, 20(2):115–121.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada. Association for Computational Linguistics.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2020. Corpus wide argument mining—a working solution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7683–7691.

M. Fromm, E. Faerman, and T. Seidl. 2019. Tacam: Topic and context aware argument mining. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 99–106, Los Alamitos, CA, USA. IEEE Computer Society.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media. In *Artificial Intelligence: Methods and Applications*, pages 287–299, Cham. Springer International Publishing.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Angelos Katharopoulos and François Fleuret. 2018. Not all samples are created equal: Deep learning with importance sampling. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2530–2539. PMLR.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Stefan Larson, Anish Mahendran, Andrew Lee, Jonathan K. Kummerfeld, Parker Hill, Michael A. Laurenzano, Johann Hauswald, Lingjia Tang, and Jason Mars. 2019. Outlier detection for improved data quality and diversity in dialog systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 517–527, Minneapolis, Minnesota. Association for Computational Linguistics.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia potentia Est—On the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 1950–1965. Curran Associates, Inc.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *ArXiv preprint*, abs/1907.11692, Version 1.

Kai Lv, Yuqing Yang, Tengxiao Liu, Qipeng Guo, and Xipeng Qiu. 2024. Full parameter fine-tuning for large language models with limited resources. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8187–8198, Bangkok, Thailand. Association for Computational Linguistics.

Huy Nguyen and Diane Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1137, Berlin, Germany. Association for Computational Linguistics.

OpenAI. 2023. ChatGPT (September 35 Version) [Large Language Model]. `https://chat.openai.com/chat`.

OpenAI. 2024. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774, Version 6.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

Gil Rocha, Henrique Lopes Cardoso, Jonas Belouadi, and Steffen Eger. 2023. Cross-genre argument mining: Can language models automatically fill in missing discourse markers? *ArXiv preprint*, abs/2306.04314, Version 1.

Andreas Rücklé, Jonas Pfeiffer, and Iryna Gurevych. 2020. MultiCQA: Zero-shot transfer of self-supervised text matching models on a massive scale. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2471–2486, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *KI - Künstliche Intelligenz*, 35(3):329–341.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656, Online. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8968–8975.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971, Version 1.

Michiel van der Meer, Myrthe Reuver, Urja Khurana, Lea Krause, and Selene Baez Santamaria. 2022. Will it blend? mixing training paradigms & prompting for argument quality prediction. In *Proceedings of the 9th Workshop on Argument Mining*, pages 95–103, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29, pages 3630–3638. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32, pages 3261–3275. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Vikas Yadav, Hyuk joon Kwon, Vijay Srinivasan, and Hongxia Jin. 2024. Explicit over implict: Explicit diversity conditions for effective question answer generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6876–6882, Torino, Italia. ELRA and ICCL.

A Reproducibility Criteria

A.1 Dataset

The new dataset FS150T-Corpus consists of 21,600 samples over 150 controversial topics with 144 samples each. We index the CommonCrawl[6] dump *CC-MAIN-2016-07* via ElasticSearch[7] and use all 150 controversial topics to search and extract texts for the crowdsourcing process. To split the texts into sentences, we use NLTK 3.7 (Bird et al., 2009). The topics for the dataset are a collection of controversially discussed subjects from the domains of, amongst others, politics, technology, and economy. They were gathered manually from twitter and reddit trends, as well as various discussion forums. See Table 7 for a list of all topics in alphabetical order, including the semantically closest topic for each given in cosine similarity. We compute the similarity of two topics by averaging the pairwise cosine similarity of all sentences for two topics. The embeddings were generated by using the sentence-transformer library[8] (version 2.2.2) (Reimers and Gurevych, 2019). As the highest cosine similarity between two topics is only 0.25, this indicates low overlapping in general.

We split the dataset into a *train*, *development*, and *test* set. There is no overlap between topics of the sets or with topics of the UKP Corpus (see Table 8). The dataset language is English and the annotation guidelines for the crowdsourcing process are taken from Stab et al. (2018). See Tables 1 and 2 for more statistics and examples about the dataset.

The crowdsourcing costs on Amazon Mechanical Turk[9] amount to a total of $3,266. The study was open to all people located in the US and we paid well above the US federal minimum wage of $7.25/hour. Each sample was annotated by seven independent, anonymous annotators. We asked the annotators to label each sample they were presented with (consisting of the guidelines, a topic, and a respective sentence) into categories *pro*, *contra*, or *none*. We design our guidelines based on Stab et al. (2018), i.e. a sentence is only to be labeled as *pro* or *contra* argument, if it holds evidence for why the sentence supports or opposes the topic. If the sentence holds no such evidence or is unrelated to the topic, it is labeled as no argument (*none*). To generate gold labels, we apply the MACE denoising tool (Hovy et al., 2013) with a threshold of 0.9 as done in (Stab et al., 2018). Finally, two experts were asked to annotate 100 randomly picked samples from the dataset. We create gold labels in the same way as for the crowdworker annotations. The Cohen's κ between expert and crowdworkers is .52, which can be interpreted as "moderate" agreement (Landis and Koch, 1977) and reasonable for the complexity of the task and the large amount of different and difficult to understand topics.

A.2 Models

A.2.1 ERNIE 2.0, FLAN-T5 XL

We tune both models on the full training sets with all combinations of four different learning rates ($1 * 10^{-4}$, $1 * 10^{-5}$, $3 * 10^{-5}$, $5 * 10^{-5}$) and three batch sizes (4, 8, 16). All models are trained over 5 epochs and we use the best model (always determined on the development set by highest F_1 macro)

[6] https://commoncrawl.org
[7] https://www.elastic.co
[8] https://github.com/UKPLab/sentence-transformers
[9] https://www.mturk.com/

	ERNIE 2.0		FLAN-T5 XL	
	Learning Rate	Batch Size	Learning Rate	Batch Size
FS150T-Corpus	$1 * 10^{-5}$	4	$1 * 10^{-4}$	4
IAM-Corpus	$1 * 10^{-5}$	16	$1 * 10^{-4}$	4
IBM-Corpus	$1 * 10^{-5}$	8	$1 * 10^{-4}$	16

Table 5: Best hyperparameters for ERNIE 2.0 and FLAN-T5 XL on all datasets.

to fix the hyperparameters for the actual experiments. Due to unstable performance on low sample sizes, we decide to always train on 6 different seeds (for tuning and actual experiments), but only leverage the averaged results on the best 3 of them. To better understand the reasoning behind this approach, we show the difference of using the best 3 seeds or all 6 seeds on the topic experiments for the FS150T-Corpus with ERNIE 2.0 (see Figure 6). As can be seen, using 6 seeds shows an overall higher standard deviation and lower performance, especially for lower sample sizes. The final hyperparameters found for each dataset and model are listed in Table 5. Training ERNIE 2.0 (110M parameters) takes approx. 25 minutes on a single NVIDIA P-100 (one seed) with the full training set of the FS150T-Corpus and approx. 45 minutes for our FLAN-T5 XL encoder (1.3B parameters) on a single NVIDIA A10 with LoRa rank 16. We use PyTorch 2.1.0 (Paszke et al., 2019) and transformers 4.37.1 (Wolf et al., 2020) to run the models and scikit-learn 0.23.2 (Pedregosa et al., 2011) to compute the metrics.

A.2.2 Llama2-70B-Chat, ChatGPT

We use Llama2-70B-Chat (Touvron et al., 2023) in a 4bit quantized version[10] that we run on four NVIDIA A6000 with vLLM (Kwon et al., 2023) and ChatGPT (gpt-3.5-turbo, September 25 Version) (OpenAI, 2023) via the OpenAI API[11]. We defined prompts that closely resemble the definitions for the respective dataset (for FS150T-Corpus, the definition from Stab et al. (2018) is used) and list all of them in Table 6.

B Dataset Licenses

We provide a list of all used datasets with their licenses:

- FS150T-Corpus (ours): Only annotations are included and licensed under CC-BY-SA 4.0. The annotated texts have to be extracted via script[2] from CommonCrawl (see also A.1) or requested[1].
- IAM-Corpus (Cheng et al., 2022): The authors do not provide a license. The data is extracted from English Wikipedia.
- UKP Corpus (Stab et al., 2018): Licensed under CC-BY-NC.
- IBM-Corpus (Ein-Dor et al., 2020): Licensed under CC-BY-SA 3.0.

C Additional Figures

This sections holds figures that include information about the standard deviation of sample (see Figures 5) and topic experiments (see Figures 7) that is left out in the main paper for better readability (see Sections 6.1 and 6.2).

[10] https://huggingface.co/TheBloke/Llama-2-70B-chat-AWQ/commit/ad4d622cb488138748dd28a0ca95c2b34dbe3964

[11] https://platform.openai.com/docs/api-reference

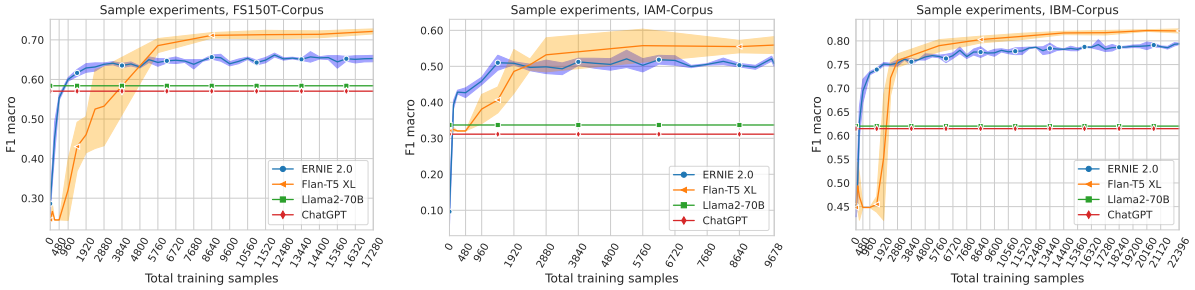

Figure 5: Sample experiments for FS150T-/IAM- and IBM-Corpus on ERNIE 2.0, FLAN-T5 XL, Llama2-70B, and ChatGPT in F_1 macro and with standard deviation.

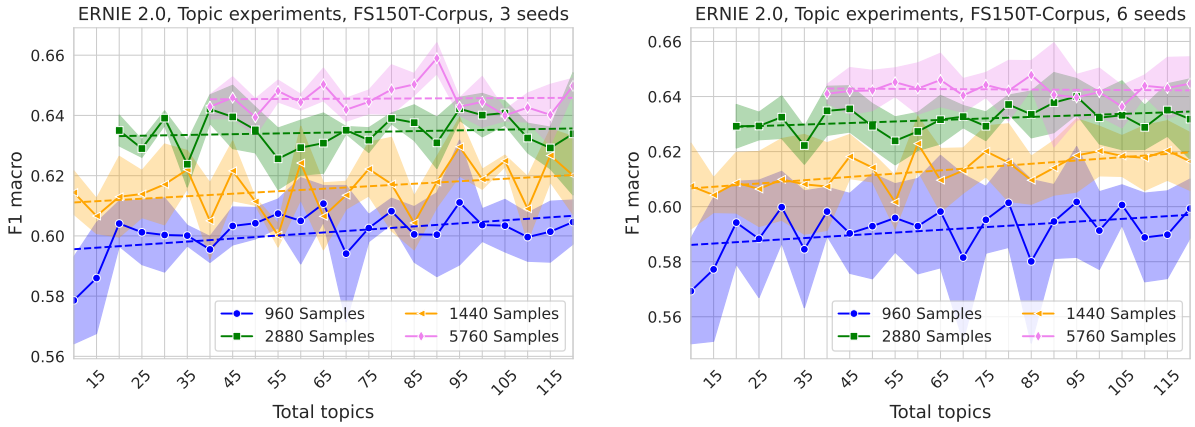

Figure 6: Same topic experiments with ERNIE 2.0 on FS150T-Corpus, but taking only the best 3 seeds on the development set (left figure) or taking all 6 seeds (right side) into account. Using 6 seeds shows higher standard deviation and lower performance, especially for smaller sample sizes.

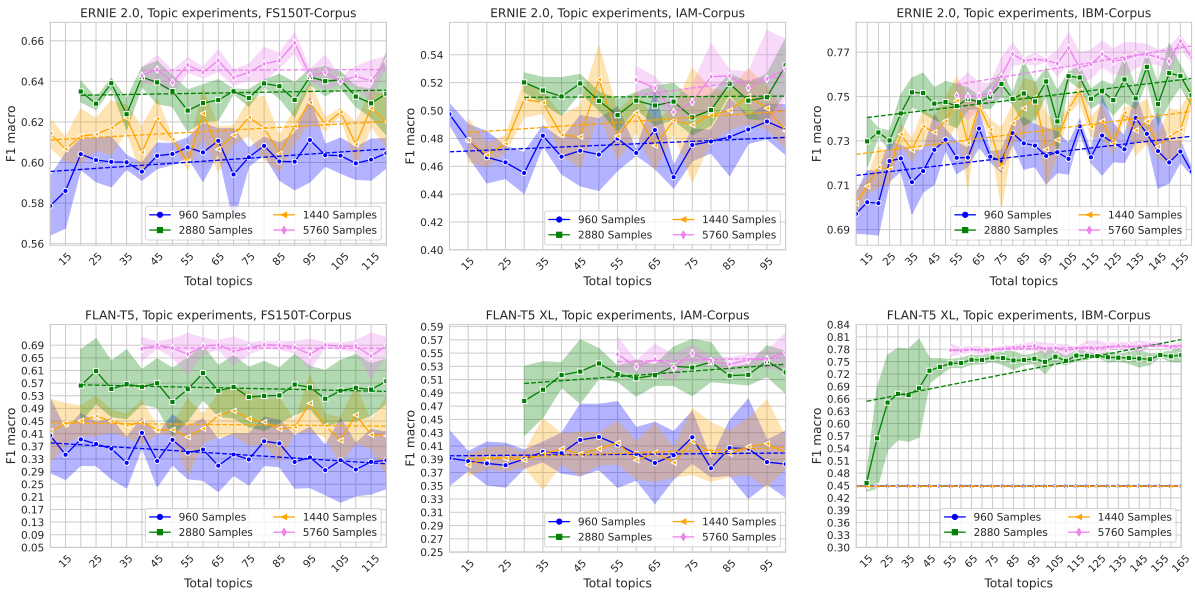

Figure 7: Topic experiments for FS150T-/IAM- and IBM-Corpus on ERNIE 2.0 and FLAN-T5 XL in F_1 macro and with standard deviation.

Dataset	Prompt 1	Prompt 2	Prompt 3
FS150T-Corpus	Decide if the below sentence is an argument with regard to the given topic. We define an argument as a span of text expressing evidence or reasoning that can be used to either support or oppose a given topic. An argument need not be “direct” or self-contained—it may pre-suppose some common or domain knowledge, or the application of commonsense reasoning—but it must be unambiguous in its orientation to the topic. If it is no argument, label it neutral. If it is an argument, decide whether it is in favor or against the topic and label it with favor or against. Only answer with neutral, favor, or against. Sentence: [A sentence from the test set] Topic: [A respective topic from the test set] Label:	Decide if the below sentence is a pro argument (label it “pro") or a contra argument (label it “contra") regarding the given topic. If it is no argument regarding topic or no argument at all, label it "none": Sentence: [A sentence from the test set] Topic: [A respective topic from the test set] Label:	Label the sentence “[A sentence from the test set]” with “pro” if it is an argument regarding topic “[A respective topic from the test set]” or “contra” if it is an argument against the topic. Label it “none” if the sentence is no argument regarding the topic or no argument at all. Label:
IAM-Corpus	What is the stance of the following sentence regarding the given topic? Only answer with one word; “other” if it is not a claim or unrelated to the topic, “support” only if it is a claim that supports the topic, or “contest” only if it is a claim that contests the topic. Sentence: [A sentence from the test set] Topic: [A respective topic from the test set] Label:	Decide the stance of the sentence below regarding the given topic. Unrelated sentences or non-claims are labelled with “none”, supporting or contesting claims are labelled with “support” or “contest”. Sentence: [A sentence from the test set] Topic: [A respective topic from the test set] Label:	Label the sentence “[A sentence from the test set]” with “support” if it is a claim that supports topic “[A respective topic from the test set]” or with “contest” if it is a claim that contests the topic. Label it "none" if the sentence is no claim regarding the topic or unrelated to it. Label:
IBM-Corpus	Decide if the following sentence is a valid evidence with regard to the given claim. We define evidence as a sentence that clearly supports or contests the claim and is not merely a belief or a claim itself. Rather, an evidence provides an indication whether a claim is true. Only answer with one word; “valid” if the sentence is an evidence with regard to the claim, otherwise return “invalid”. Sentence: [A sentence from the test set] Claim: [A respective claim from the test set] Label:	Decide if the below sentence is a valid evidence regarding the given claim (label it “yes”) or not (label it “no”): Sentence: [A sentence from the test set] Claim: [A respective claim from the test set] Label:	Label the sentence “[A sentence from the test set]” with "valid" if it is a valid evidence for the claim “[A respective claim from the test set]” or "invalid" if not. Label:

Table 6: All prompts used for experiments with LLama2-70B and ChatGPT on FS150T-Corpus (best: Prompt 2), IAM-Corpus (best: Prompt 2), and IBM-Corpus (best: Prompt 1).

Topic	Most similar topic	Cosine sim.	Topic	Most similar topic	Cosine sim.	Topic	Most similar topic	Cosine sim.
3d printer	holography	0.11	foreign aid	us intervention	0.17	prescription drug ads	big pharma	0.19
alcohol advertising	lower drinking age	0.22	fracking	offshore drilling	0.21	progressive tax	farm subsidies	0.18
alternative medicine	big pharma	0.17	free market	progressive tax	0.16	racial profiling	reverse discrimination	0.18
amazon	ebooks	0.13	freedom of speech	usa patriot act	0.15	religious holidays	atheism	0.12
anarchism	isolationism	0.16	fuel tax	progressive tax	0.17	renewable energy	wind energy	0.20
animal dissection	animal testing	0.20	gambling	legalized prostitution	0.14	reparations for slavery	white supremacy	0.20
animal testing	animal dissection	0.20	gay marriage	gay rights	0.23	reverse discrimination	white supremacy	0.19
antibiotic usage	alternative medicine	0.16	gay rights	gay marriage	0.23	right to health care	obamacare	0.21
artificial intelligence	autonomous cars	0.12	geothermal energy	hydroelectricity	0.19	robots	autonomous cars	0.12
assisted suicide	lethal injection	0.22	global warming	man-made greenhouse gases	0.21	sanctuary cities	illegal immigration	0.20
atheism	existence of god	0.21	glyphosate	gmos	0.18	school vouchers	charter schools	0.24
autonomous cars	lower speed limit	0.19	gmos	biofuels	0.18	sex education in school	birth control	0.19
beauty contest	feminism	0.14	government surveillance	usa patriot act	0.20	sex offender registry	mandatory sentencing	0.18
big pharma	prescription drug ads	0.19	guantanamo bay detention camp	drone strikes	0.15	smart home	smartwatch	0.12
bilingual education	standardized testing	0.17	holography	3d printer	0.11	smartwatch	amazon	0.12
biofuels	offshore drilling	0.19	homeschooling	charter schools	0.18	social media	net neutrality	0.13
birth control	sex education in school	0.19	homework	homeschooling	0.17	solar energy	renewable energy	0.19
boarding school	charter schools	0.17	hydroelectricity	renewable energy	0.20	spanking	corporal punishment	0.18
border security	illegal immigration	0.21	illegal immigration	border security	0.21	sperm donor	surrogacy	0.20
brexit	foreign aid	0.14	insanity defense	mandatory sentencing	0.20	standardized testing	teacher tenure	0.18
bullfighting	factory farming	0.13	insider trading	labor unions	0.13	stem cell research	organ donation	0.18
cell phone radiation	stem cell research	0.13	isolationism	nuclear disarmament	0.19	surrogacy	sperm donor	0.20
censorship	net neutrality	0.16	jury duty	mandatory sentencing	0.17	suv	autonomous cars	0.16
charter schools	school vouchers	0.24	labor unions	unemployment insurance	0.18	teacher tenure	charter schools	0.22
cheerleading	beauty contest	0.11	legalized prostitution	monogamy	0.18	term limit	electoral college	0.18
clerical celibacy	monogamy	0.18	lethal injection	assisted suicide	0.22	tobacco advertising	electronic cigarettes	0.20
coal mining	fracking	0.15	libertarianism	right to health care	0.16	transgender rights	gay rights	0.20
community service	school vouchers	0.12	life extension	stem cell research	0.14	two-state solution	nuclear disarmament	0.18
compulsory voting	electoral college	0.25	lobbying	two-state solution	0.14	unemployment insurance	labor unions	0.18
concealed handguns	mandatory sentencing	0.15	lottery	crowdfunding	0.13	urban agriculture	farm subsidies	0.17
corporal punishment	mandatory sentencing	0.19	lower drinking age	alcohol advertising	0.22	urbanization	urban agriculture	0.12
crowdfunding	farm subsidies	0.13	lower speed limit	autonomous cars	0.19	us intervention	war on terrorism	0.19
cultured meat	factory farming	0.20	man-made greenhouse gases	global warming	0.21	usa patriot act	government surveillance	0.20
daycare	homeschooling	0.14	mandatory national service	right to health care	0.15	vaccination	animal testing	0.15
daylight saving time	solar energy	0.11	mandatory sentencing	insanity defense	0.20	vegetarianism	cultured meat	0.20
direct democracy	compulsory voting	0.22	monarchy	direct democracy	0.16	video games and violence	corporal punishment	0.14
drone strikes	war on terrorism	0.20	monogamy	gay marriage	0.23	virtual reality	artificial intelligence	0.11
ebooks	amazon	0.13	multiculturalism	white supremacy	0.17	voting machines	compulsory voting	0.24
ecotourism	urban agriculture	0.16	net neutrality	censorship	0.16	war on drugs	legalized prostitution	0.17
electoral college	compulsory voting	0.25	nuclear disarmament	isolationism	0.19	war on obesity	right to health care	0.14
electronic cigarettes	tobacco advertising	0.20	obamacare	right to health care	0.21	war on terrorism	drone strikes	0.20
executive order	obamacare	0.18	occupy wall street	white supremacy	0.14	water privatization	fracking	0.15
existence of god	atheism	0.21	offshore drilling	fracking	0.21	weather modification	man-made greenhouse gases	0.16
extraterrestrial life	existence of god	0.14	online dating service	monogamy	0.10	whaling	cultured meat	0.16
extreme sport	autonomous cars	0.09	organ donation	assisted suicide	0.19	white supremacy	reparations for slavery	0.20
factory farming	cultured meat	0.20	organic food	vegetarianism	0.19	wikileaks	government surveillance	0.17
farm subsidies	progressive tax	0.18	outsourcing	crowdfunding	0.13	wind energy	renewable energy	0.20
fast food	vegetarianism	0.17	pedelec	autonomous cars	0.14	wiretapping	government surveillance	0.16
felon voting	compulsory voting	0.23	plastic surgery	alternative medicine	0.13	women in the military	feminism	0.18
feminism	monogamy	0.19	police body cameras	government surveillance	0.15	year-round school	charter schools	0.20

Table 7: List of all 150 topics for the FS150T-Corpus, including their semantically closest topic computed via embeddings with model "all-MiniLM-L6-v2" (Reimers and Gurevych, 2019). Highest cosine similarity computed: 0.25.

Topic (FS150T-Corpus)	Most similar topic (UKP Corpus)	Cosine sim.	Topic (FS150T-Corpus)	Most similar topic (UKP Corpus)	Cosine sim.	Topic (FS150T-Corpus)	Most similar topic (UKP Corpus)	Cosine sim.
3d printer	minimum wage	0.07	fast food	minimum wage	0.08	online dating service	marijuana legalization	0.04
birth control	abortion	0.17	felon voting	death penalty	0.15	organ donation	cloning	0.16
alcohol advertising	marijuana legalization	0.15	feminism	abortion	0.17	organic food	marijuana legalization	0.08
alternative medicine	marijuana legalization	0.12	foreign aid	minimum wage	0.12	outsourcing	minimum wage	0.12
amazon	minimum wage	0.05	fracking	nuclear energy	0.14	pedelec	nuclear energy	0.07
anarchism	gun control	0.11	free market	minimum wage	0.14	plastic surgery	cloning	0.10
animal dissection	cloning	0.12	freedom of speech	gun control	0.13	police body cameras	gun control	0.12
animal testing	cloning	0.15	fuel tax	minimum wage	0.12	prescription drug ads	marijuana legalization	0.14
antibiotic usage	marijuana legalization	0.09	gambling	marijuana legalization	0.12	progressive tax	minimum wage	0.16
artificial intelligence	cloning	0.10	gay marriage	abortion	0.14	racial profiling	gun control	0.13
assisted suicide	death penalty	0.17	gay rights	abortion	0.14	religious holidays	school uniforms	0.09
atheism	cloning	0.12	geothermal energy	nuclear energy	0.15	renewable energy	nuclear energy	0.15
autonomous cars	gun control	0.09	global warming	nuclear energy	0.14	reparations for slavery	death penalty	0.13
beauty contest	school uniforms	0.10	glyphosate	marijuana legalization	0.11	reverse discrimination	school uniforms	0.12
big pharma	marijuana legalization	0.14	gmos	cloning	0.14	right to health care	abortion	0.14
bilingual education	school uniforms	0.12	government surveillance	gun control	0.12	robots	cloning	0.09
biofuels	nuclear energy	0.14	guantanamo bay detention camp	death penalty	0.12	sanctuary cities	marijuana legalization	0.12
birth control	abortion	0.17	concealed handguns	gun control	0.20	school vouchers	school uniforms	0.15
boarding school	school uniforms	0.12	holography	cloning	0.05	school vouchers	school uniforms	0.15
border security	marijuana legalization	0.12	homeschooling	school uniforms	0.13	sex education in school	abortion	0.17
brexit	minimum wage	0.11	homework	school uniforms	0.13	sex offender registry	death penalty	0.14
bullfighting	gun control	0.08	hydroelectricity	nuclear energy	0.15	smart home	nuclear energy	0.07
cell phone radiation	cloning	0.10	illegal immigration	marijuana legalization	0.12	smartwatch	minimum wage	0.04
censorship	gun control	0.11	insanity defense	death penalty	0.17	social media	school uniforms	0.07
charter schools	school uniforms	0.15	insider trading	minimum wage	0.11	solar energy	nuclear energy	0.14
cheerleading	school uniforms	0.10	isolationism	gun control	0.12	spanking	death penalty	0.09
clerical celibacy	abortion	0.12	jury duty	death penalty	0.14	sperm donor	cloning	0.17
stem cell research	cloning	0.21	labor unions	minimum wage	0.16	standardized testing	school uniforms	0.14
coal mining	nuclear energy	0.11	legalized prostitution	marijuana legalization	0.15	stem cell research	cloning	0.21
community service	minimum wage	0.09	lethal injection	death penalty	0.22	surrogacy	abortion	0.17
compulsory voting	minimum wage	0.12	libertarianism	abortion	0.13	suv	nuclear energy	0.05
concealed handguns	gun control	0.20	life extension	cloning	0.12	teacher tenure	school uniforms	0.15
corporal punishment	death penalty	0.16	lobbying	gun control	0.13	term limit	gun control	0.12
crowdfunding	minimum wage	0.12	lottery	minimum wage	0.10	tobacco advertising	marijuana legalization	0.17
cultured meat	cloning	0.12	lower drinking age	marijuana legalization	0.15	transgender rights	abortion	0.13
daycare	school uniforms	0.08	lower speed limit	gun control	0.09	two-state solution	gun control	0.10
daylight saving time	minimum wage	0.09	man-made greenhouse gases	nuclear energy	0.13	unemployment insurance	minimum wage	0.18
lethal injection	death penalty	0.22	mandatory national service	gun control	0.12	urban agriculture	minimum wage	0.10
direct democracy	gun control	0.12	mandatory sentencing	death penalty	0.19	urbanization	minimum wage	0.10
drone strikes	gun control	0.14	war on drugs	marijuana legalization	0.19	us intervention	gun control	0.12
ebooks	minimum wage	0.05	unemployment insurance	minimum wage	0.18	usa patriot act	gun control	0.15
ecotourism	nuclear energy	0.10	monarchy	abortion	0.09	vaccination	cloning	0.12
electoral college	gun control	0.10	monogamy	abortion	0.13	vegetarianism	abortion	0.05
electronic cigarettes	marijuana legalization	0.13	multiculturalism	school uniforms	0.10	video games and violence	gun control	0.12
executive order	abortion	0.15	net neutrality	minimum wage	0.10	virtual reality	cloning	0.07
existence of god	cloning	0.12	nuclear disarmament	nuclear energy	0.18	voting machines	gun control	0.09
extraterrestrial life	cloning	0.12	nuclear disarmament	nuclear energy	0.18	war on drugs	marijuana legalization	0.19
extreme sport	school uniforms	0.07	obamacare	minimum wage	0.14	war on obesity	marijuana legalization	0.11
factory farming	cloning	0.12	occupy wall street	gun control	0.11	war on terrorism	gun control	0.13
farm subsidies	minimum wage	0.15	offshore drilling	nuclear energy	0.14	water privatization	nuclear energy	0.11

Table 8: List of all 150 topics for the FS150T-Corpus and their semantically closest topic from the UKP Corpus, computed via embeddings with model "all-MiniLM-L6-v2" (Reimers and Gurevych, 2019). Highest cosine similarity computed: 0.22.