Topic-Guided Sampling For Data-Efficient Multi-Domain Stance Detection

Erik Arakelyan[1], **Arnav Arora**[2], **Isabelle Augenstein**[3]
Department of Computer Science
University of Copenhagen
Copenhagen Denmark
{erik.a,aar,augenstein}@di.ku.dk

Abstract

Stance Detection is concerned with identifying the attitudes expressed by an author towards a target of interest. This task spans a variety of domains ranging from social media opinion identification to detecting the stance for a legal claim. However, the framing of the task varies within these domains, in terms of the data collection protocol, the label dictionary and the number of available annotations. Furthermore, these stance annotations are significantly imbalanced on a per-topic and inter-topic basis. These make multi-domain stance detection a challenging task, requiring standardization and domain adaptation. To overcome this challenge, we propose **T**opic **E**fficient **St**anc**E** **D**etection (TESTED), consisting of a topic-guided diversity sampling technique and a contrastive objective that is used for fine-tuning a stance classifier. We evaluate the method on an existing benchmark of 16 datasets with in-domain, i.e. all topics seen and out-of-domain, i.e. unseen topics, experiments. The results show that our method outperforms the state-of-the-art with an average of 3.5 F1 points increase in-domain, and is more generalizable with an averaged increase of 10.2 F1 on out-of-domain evaluation while using $\leq 10\%$ of the training data. We show that our sampling technique mitigates both inter- and per-topic class imbalances. Finally, our analysis demonstrates that the contrastive learning objective allows the model a more pronounced segmentation of samples with varying labels.

1 Introduction

The goal of stance detection is to identify the viewpoint expressed by an author within a piece of text towards a designated topic (Mohammad et al., 2016). Such analyses can be used in a variety of domains ranging from identifying claims within political or ideological debates (Somasundaran and Wiebe, 2010; Thomas et al., 2006), identifying mis- and disinformation (Hanselowski et al., 2018; Hardalov et al., 2022a), public health policymaking (Glandt et al., 2021; Hossain et al., 2020; Osnabrügge et al., 2023), news recommendation (Reuver et al., 2021) to investigating attitudes voiced on social media (Qazvinian et al., 2011; Augenstein et al., 2016; Conforti et al., 2020). However, in most domains, and even more so for cross-domain stance detection, the exact formalisation of the task gets blurry, with varying label sets and their corresponding definitions, data collection protocols and available annotations. Furthermore, this is accompanied by significant changes in the topic-specific vocabulary (Somasundaran and Wiebe, 2010; Wei and Mao, 2019), text style (Pomerleau and Rao, 2017; Ferreira and Vlachos, 2016) and topics mentioned either explicitly (Qazvinian et al., 2011; Walker et al., 2012) or implicitly (Hasan and Ng, 2013; Derczynski et al., 2017). Recently, a benchmark of 16 datasets (Hardalov et al., 2021) covering a variety of domains and topics has been proposed for testing stance detection models across multiple domains. It must be noted that these datasets are highly imbalanced, with an imbalanced label distribution between the covered topics, i.e. inter-topic and within each topic, i.e. per-topic, as can be seen in Figure 2 and Figure 3. This further complicates the creation of a robust stance detection classifier.

Given the inherent skew present within the dataset and variances within each domain, we propose a topic-guided diversity sampling method, which produces a data-efficient representative subset while mitigating label imbalances. These samples are used for fine-tuning a Pre-trained Language Model (PLM), using a contrastive learning objective to create a robust stance detection model. These two components form our **T**opic **E**fficient **St**anc**E** **D**etection (TESTED) framework, as seen in Figure 1, and are analysed separately to pinpoint the factors impacting model performance and robustness. We test our method on

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 13448–13464
July 9-14, 2023 ©2023 Association for Computational Linguistics

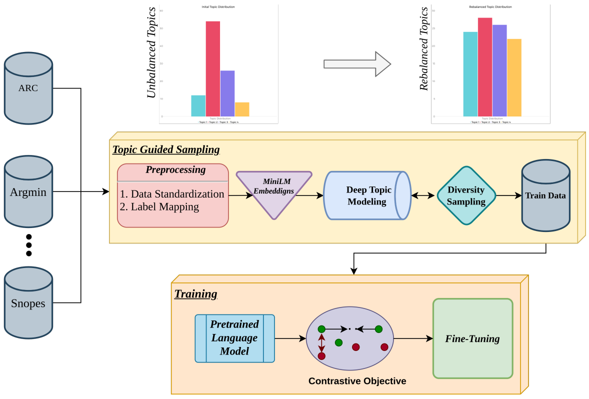

Figure 1: The two components of TESTED: Topic Guided Sampling (top) and training with contrastive objective (bottom).

the multi-domain stance detection benchmark by Hardalov et al. (2021), achieving state-of-the-art results with both in-domain, i.e. all topics seen and out-of-domain, i.e. unseen topics evaluations. Note though that TESTED could be applied to any text classification setting.

In summary, our **contributions** are:

- We propose a novel framework (TESTED) for predicting stances across various domains, with data-efficient sampling and contrastive learning objective;
- Our proposed method achieves SOTA results both in-domain and out-of-domain;
- Our analysis shows that our topic-guided sampling method mitigates dataset imbalances while accounting for better performance than other sampling techniques;
- The analysis shows that the contrastive learning objective boosts the ability of the classifier to differentiate varying topics and stances.

2 Related Work

Stance Detection is an NLP task which aims to identify an author's attitude towards a particular topic or claim. The task has been widely explored in the context of mis- and disinformation detection (Ferreira and Vlachos, 2016; Hanselowski et al., 2018; Zubiaga et al., 2018b; Hardalov et al., 2022a), sentiment analysis (Mohammad et al., 2017; Aldayel and Magdy, 2019) and argument mining (Boltužić and Šnajder, 2014; Sobhani et al., 2015; Wang et al., 2019). Most papers formally define stance detection as a pairwise sequence classification where stance targets are provided (Küçük and Can, 2020). However, with the emergence of different data sources, ranging from debating platforms (Somasundaran and Wiebe, 2010; Hasan and Ng, 2014; Aharoni et al., 2014) to social media (Mohammad et al., 2016; Derczynski et al., 2017), and new applications (Zubiaga et al., 2018a; Hardalov et al., 2022a), this formal definition has been subject to variations w.r.t. the label dictionary inferred for the task.

Previous research has predominantly focused on a specific dataset or domain of interest, outside of a few exceptions like multi-target (Sobhani et al., 2017; Wei et al., 2018) and cross-lingual (Hardalov et al., 2022b) stance detection. In contrast, our work focuses on multi-domain stance detection, while evaluating in- and out-of-domain on a 16 dataset benchmark with state-of-the-art baselines (Hardalov et al., 2021).

Topic Sampling Our line of research is closely associated with diversity (Ren et al., 2021) and importance (Beygelzimer et al., 2009) sampling and their applications in natural language processing (Zhu et al., 2008; Zhou and Lampouras, 2021). Clustering-based sampling approaches have been used for automatic speech recognition (Syed et al., 2016), image classification (Ranganathan et al., 2017; Yan et al., 2022) and semi-supervised active learning (Buchert et al., 2022) with limited use for textual data (Yang et al., 2014) through topic modelling (Blei et al., 2001). This research proposes an importance-weighted topic-guided diversity sampling method that utilises deep topic models, for mitigating inherent imbalances present in the data, while preserving relevant examples.

Contrastive Learning has been used for tasks where the expected feature representations should be able to differentiate between similar and divergent inputs (Liu et al., 2021; Rethmeier and Augenstein, 2023). Such methods have been used for image classification (Khosla et al., 2020), captioning (Dai and Lin, 2017) and textual representations (Giorgi et al., 2021; Jaiswal et al., 2020; Ostendorff et al., 2022). The diversity of topics (Qazvinian et al., 2011; Walker et al., 2012; Hasan and Ng, 2013), vocabulary (Somasundaran and Wiebe, 2010; Wei and Mao, 2019) and expression styles (Pomerleau and Rao, 2017) common for stance detection can be tackled with contrastive objectives, as seen for similar sentence embedding and classification tasks (Gao et al., 2021; Yan et al., 2021).

3 Datasets

Our study uses an existing multi-domain dataset benchmark (Hardalov et al., 2021), consisting of 16 individual datasets split into four source groups: *Debates, News, Social Media, Various*. The categories include datasets about debating and political claims including arc (Hanselowski et al., 2018; Habernal et al., 2018), iac1 (Walker et al., 2012), perspectum (Chen et al., 2019), poldeb (Somasundaran and Wiebe, 2010), scd (Hasan and Ng, 2013), news like emergent (Ferreira and Vlachos, 2016), fnc1 (Pomerleau and Rao, 2017), snopes (Hanselowski et al., 2019), social media like mtsd (Sobhani et al., 2017), rumour (Qazvinian et al., 2011), semeval2016t6 (Mohammad et al., 2016), semeval2019t7 (Derczynski et al., 2017), wtwt (Conforti et al., 2020) and datasets that cover a variety of diverse topics like argmin (Stab et al., 2018), ibmcs (Bar-Haim et al., 2017) and vast (Allaway and McKeown, 2020). Overall statistics for all of the datasets can be seen in Appendix C.

3.1 Data Standardisation

As the above-mentioned stance datasets from different domains possess different label inventories, the stance detection benchmark by Hardalov et al. (2021) introduce a mapping strategy to make the class inventory homogeneous. We adopt that same mapping for a fair comparison with prior work, shown in Appendix C.

4 Methods

Our goal is to create a stance detection method that performs strongly on the topics known during training and can generalize to unseen topics. The benchmark by Hardalov et al. (2021) consisting of 16 datasets is highly imbalanced w.r.t the inter-topic frequency and per-topic label distribution, as seen in Figure 2.

These limitations necessitate a novel experimental pipeline. The first component of the pipeline we propose is an importance-weighted topic-guided diversity sampling method that allows the creation of supervised training sets while mitigating the inherent imbalances in the data. We then create a stance detection model by fine-tuning a Pre-trained Language Model (PLM) using a contrastive objective.

4.1 Topic-Efficient Sampling

We follow the setting in prior work on data-efficient sampling (Buchert et al., 2022; Yan et al., 2022), framing the task as a selection process between multi-domain examples w.r.t the theme discussed within the text and its stance. This means that given a set of datasets $\mathcal{D} = (\mathcal{D}_1, \ldots \mathcal{D}_n)$ with their designated documents $\mathcal{D}_i = (d_i^1, \ldots d_i^m)$, we wish to select a set of diverse representative examples $\mathcal{D}_{\textbf{train}}$, that are balanced w.r.t the provided topics $\mathcal{T} = (t_1, \ldots t_q)$ and stance labels $L = (l_1, \ldots l_k)$.

Diversity Sampling via Topic Modeling We thus opt for using topic modelling to produce a supervised subset from all multi-domain datasets. Selecting annotated examples during task-specific fine-tuning is a challenging task (Shao et al., 2019), explored extensively within active learning research (Hino, 2020; Konyushkova et al., 2017). Random sampling can lead to poor generalization and knowledge transfer within the novel problem domain (Das et al., 2021; Perez et al., 2021). To mitigate the inconsistency caused by choosing sub-optimal examples, we propose using deep unsupervised topic models, which allow us to sample relevant examples for each topic of interest. We further enhance the model with an importance-weighted diverse example selection process (Shao et al., 2019; Yang et al., 2015) within the relevant examples generated by the topic model. The diversity maximisation sampling is modeled similarly to Yang et al. (2015).

The topic model we train is based on the technique proposed by Angelov (2020) that tries to find topic vectors while jointly learning document and word semantic embeddings. The topic model is initialized with weights from the *all-MiniLM-L6* PLM, which has a strong performance on sentence embedding benchmarks (Wang et al., 2020). It is shown that learning unsupervised topics in this fashion maximizes the total information gained, about all texts $\mathcal{D}$ when described by all words $\mathcal{W}$.

$$\mathcal{I}(\mathcal{D}, \mathcal{W}) = \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} P(d, w) \log \left(\frac{P(d, w)}{P(d)P(w)} \right)$$

This characteristic is handy for finding relevant samples across varying topics, allowing us to search within the learned documents d_i. We train a deep topic model $\mathcal{M}_{topic}$ using multi-domain data $\mathcal{D}$ and obtain topic clusters $\mathcal{C} = (\mathcal{C}_i, \ldots \mathcal{C}_t)$,

Algorithm 1 Topic Efficient Sampling

Require: $S \geq 0$ ▷ Sampling Threshold
Require: $Avg \in \{moving, exp\}$
Ensure: $|\mathcal{C}| > 0$
  $\mathcal{D}_{\textbf{train}} \leftarrow \{\}$
  $I \leftarrow \{\frac{|\mathcal{C}_1|}{\sum_{\mathcal{C}_i \in \mathcal{C}} \mathcal{C}_i} \dots \frac{|\mathcal{C}_t|}{\sum_{\mathcal{C}_i \in \mathcal{C}} \mathcal{C}_i}\}$ ▷ Cluster Importances
  for $\mathcal{C}_i \in \mathcal{C}$ **do** ▷ Iterating for each cluster
    $\mathcal{E}_i \leftarrow \{PLM(d_i^1) \dots\} = \{\mathbf{e}_i^1 \dots \mathbf{e}_i^m\}$
    $s_i \leftarrow max(1, S \cdot I_i)$ ▷ Threshold per cluster
    $j \leftarrow 0$
    $\boldsymbol{cent}_0 \leftarrow \frac{\sum_{\mathbf{e}_i \in \mathcal{E}} \mathbf{e}_i}{|\mathcal{E}|}$ ▷ Centroid of the cluster
    while $j \leq s_i$ **do**
      $\boldsymbol{sim} = \frac{\langle \mathcal{E}, cent \rangle}{\|\mathcal{E}\| \|cent\|}$ ▷ Similarity Ranking
      $sample = \arg\operatorname{sort}(\boldsymbol{sim}, \textbf{Ascending})[0]$
  ▷ Take the sample most diverse from the centroid
      $\mathcal{D}_{\textbf{train}} \leftarrow \mathcal{D}_{\textbf{train}} \cup sample$
      $j \leftarrow j + 1$
      $\boldsymbol{cent}_j \leftarrow \begin{cases} \alpha \cdot \mathbf{e}_{sample} + (1 - \alpha) \cdot cent_{j-1} & exp \\ \frac{(j-1)}{j} \cdot \boldsymbol{cent}_j + \frac{\mathbf{e}_{sample}}{j} & moving \end{cases}$
        ▷ Centroid update w.r.t. sampled data
    end while
  end for
  return $\mathcal{D}_{\textbf{train}}$

where $|\mathcal{C}| = t$ is the number of topic clusters. We obtain the vector representation for $\forall d_i$ from the tuned PLM embeddings $\mathcal{E} = (e_1, \dots e_m)$ in $\mathcal{M}_{topic}$, while iteratively traversing through the clusters $\mathcal{C}_i \in \mathcal{C}$.

Our sampling process selects increasingly more diverse samples after each iteration. This search within the relevant examples is presented in Algorithm 1. This algorithm selects a set of diverse samples from the given multi-domain datasets $\mathcal{D}$, using the clusters from a deep topic model $\mathcal{M}_{topic}$ and the sentence embeddings $\mathcal{E}$ of the sentences as a basis for comparison. The algorithm starts by selecting a random sentence as the first diverse sample and uses this sentence to calculate a "centroid" embedding. It then iteratively selects the next most dissimilar sentence to the current centroid, until the desired number of diverse samples is obtained.

4.2 Topic-Guided Stance Detection

Task Formalization Given the topic, t_i for each document d_i in the generated set $\mathcal{D}_{\textbf{train}}$ we aim to classify the stance expressed within that text towards the topic. For a fair comparison with prior work, we use the label mapping from the previous multi-domain benchmark (Hardalov et al., 2021) and standardise the original labels L into a five-way stance classification setting, $S = \{\text{Positive, Negative, Discuss, Other, Neutral}\}$. Stance detection can be generalized as pairwise sequence classification, where a model learns a mapping $f : (d_i, t_i) \rightarrow S$. We combine the textual sequences with the stance labels to learn this mapping. The combination is implemented using a simple prompt commonly used for NLI tasks (Lan et al., 2020; Raffel et al., 2020; Hambardzumyan et al., 2021), where the textual sequence becomes the premise and the topic the hypothesis.

[CLS] premise: *premise*
hypothesis: *topic* [EOS]

The result of this process is a supervised dataset for stance prediction $\mathcal{D}_{\textbf{train}} = ((Prompt(d_1, t_1), s_1) \dots (Prompt(d_n, t_n), s_n))$ where $\forall s_i \in S$. This method allows for data-efficient sampling, as we at most sample 10% of the data while preserving the diversity and relevance of the selected samples. The versatility of the method allows *TESTED* to be applied to any text classification setting.

Tuning with a Contrastive Objective After obtaining the multi-domain supervised training set $\mathcal{D}_{\textbf{train}}$, we decided to leverage the robustness of PLMs, based on a transformer architecture (Vaswani et al., 2017) and fine-tune on $\mathcal{D}_{\textbf{train}}$ with a single classification head. This effectively allows us to transfer the knowledge embedded within the PLM onto our problem domain. For standard fine-tuning of the stance detection model $\mathcal{M}_{stance}$ we use cross-entropy as our initial loss:

$$\mathcal{L}_{CE} = -\sum_{i \in S} y_i \log \left(\mathcal{M}_{stance}(d_i)\right) \quad (1)$$

Here y_i is the ground truth label. However, as we operate in a multi-domain setting, with variations in writing vocabulary, style and covered topics, it is necessary to train a model where similar sentences have a homogeneous representation within the embedding space while keeping contrastive pairs distant. We propose a new contrastive objective based on the *cosine* distance between the samples to accomplish this. In each training batch $B = (d_1, \dots d_b)$, we create a matrix of contrastive pairs $\mathcal{P} \in \mathcal{R}^{b \times b}$, where $\forall i, j = \overline{1, b}$, $\mathcal{P}_{ij} = 1$

if i-th and j-th examples share the same label and -1 otherwise. The matrices can be precomputed during dataset creation, thus not adding to the computational complexity of the training process. We formulate our pairwise contrastive objective $\mathcal{L}_{CL}(x_i, x_j, \mathcal{P}_{ij})$ using matrix $\mathcal{P}$.

$$\mathcal{L}_{CL} = \begin{cases} e(1 - e^{\cos(x_i,x_j)-1}), \mathcal{P}_{ij} = 1 \\ e^{\max(0,\cos(x_i,x_j)-\beta)} - 1, \mathcal{P}_{ij} = -1 \end{cases} \quad (2)$$

Here x_i, x_j are the vector representations of examples d_i, d_j. The loss is similar to cosine embedding loss and soft triplet loss (Barz and Denzler, 2020; Qian et al., 2019); however, it penalizes the opposing pairs harsher because of the exponential nature, but does not suffer from computational instability as the values are bounded in the range $[0, e - \frac{1}{e}]$. The final loss is:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{CL} \quad (3)$$

We use the fine-tuning method from Mosbach et al. (2021); Liu et al. (2019) to avoid the instability caused by catastrophic forgetting, small-sized fine-tuning datasets or optimization difficulties.

5 Experimental Setup

5.1 Evaluation

We evaluate our method on the 16 dataset multi-domain benchmark and the baselines proposed by Hardalov et al. (2021). To directly compare with prior work, we use the same set of evaluation metrics: macro averaged F1, precision, recall and accuracy.

5.2 Model Details

We explore several PLM transformer architectures within our training and classification pipelines in order to evaluate the stability of the proposed technique. We opt to finetune a pre-trained *roberta-large* architecture (Liu et al., 2019; Conneau et al., 2020). For fine-tuning, we use the method introduced by Mosbach et al. (2021), by adding a linear warmup on the initial 10% of the iteration raising the learning rate to $2e^{-5}$ and decreasing it to 0 afterwards. We use a weight decay of $\lambda = 0.01$ and train for 3 epochs with global gradient clipping on the stance detection task. We further show that learning for longer epochs does not yield sizeable improvement over the initial fine-tuning. The optimizer used for experimentation is an AdamW (Loshchilov and Hutter, 2019) with a bias correction component added to stabilise the experimentation (Mosbach et al., 2021).

Topic Efficiency Recall that we introduce a topic-guided diversity sampling method within ***TESTED***, which allows us to pick relevant samples per topic and class for further fine-tuning. We evaluate its effectiveness by fine-tuning PLMs on the examples it generates and comparing it with training on a random stratified sample of the same size.

6 Results and Analysis

In this section, we discuss and analyze our results, while comparing the performance of the method against the current state-of-the-art (Hardalov et al., 2021) and providing an analysis of the topic efficient sampling and the contrastive objective.

6.1 Stance Detection

In-domain We train on our topic-efficient subset $\mathcal{D}_{\textbf{train}}$ and test the method on all datasets $\mathcal{D}$ in the multi-domain benchmark. Our method TESTED is compared to MoLE (Hardalov et al., 2021), a strong baseline and the current state-of-the-art on the benchmark. The results, presented in Table 1, show that TESTED has the highest average performance on in-domain experiments with an increase of 3.5 F1 points over MoLE, all while using $\leq 10\%$ of the amount of training data in our subset $\mathcal{D}_{\textbf{train}}$ sampled from the whole dataset $\mathcal{D}$. Our method is able to outperform all the baselines on 10 out of 16 datasets. On the remaining 6 datasets the maximum absolute difference between TESTED and MoLE is 1.1 points in F1. We also present ablations for TESTED, by replacing the proposed sampling method with other alternatives, removing the contrastive objective or both simultaneously. Replacing Topic Efficient sampling with either *Random* or *Stratified* selections deteriorates the results for all datasets with an average decrease of 8 and 5 F1 points, respectively. We attribute this to the inability of other sampling techniques to maintain inter-topic distribution and per-topic label distributions balanced while selecting diverse samples. We further analyse how our sampling technique tackles these tasks in subsection 6.2. We also see that removing the contrastive loss also results in a deteriorated performance across all the datasets with an average decrease of 3 F1 points. In particular, we see a more significant decrease in datasets with similar topics and textual expressions, i.e. *poldeb*

	F_1 avg.	arc	iac1	perspectrum	poldeb	scd	emergent	fnc1	snopes	mtsd	rumor	semeval16	semeval19	wtwt	argmin	ibmcs	vast
Majority class baseline	27.60	21.45	21.27	34.66	39.38	35.30	21.30	20.96	43.98	19.49	25.15	24.27	22.34	15.91	33.83	34.06	17.19
Random baseline	35.19	18.50	30.66	50.06	48.67	50.08	31.83	18.64	45.49	33.15	20.43	31.11	17.02	20.01	49.94	50.08	33.25
MoLE	65.55	63.17	38.50	85.27	50.76	**65.91**	**83.74**	75.82	75.07	**65.08**	**67.24**	**70.05**	57.78	68.37	**63.73**	79.38	38.92
TESTED (Our Model)	**69.12**	**64.82**	**56.97**	**83.11**	**52.76**	64.71	82.10	**83.17**	**78.61**	63.96	66.58	69.91	**58.72**	**70.98**	62.79	**88.06**	**57.47**
Topic → Random Sampling	61.14	53.92	42.59	77.68	44.08	52.54	67.55	75.60	72.67	56.35	59.08	66.88	57.28	69.32	52.02	76.93	53.80
Topic → Stratified Sampling	64.01	50.27	51.57	77.78	46.67	62.13	79.00	77.90	76.44	61.50	64.92	68.45	51.96	69.47	56.76	78.30	51.16
- Contrastive Objective	65.63	61.11	55.50	81.85	43.81	63.04	80.84	79.05	73.43	62.18	61.57	60.17	56.06	68.79	59.51	86.94	56.35
Topic Sampling → Stratified - Contrastive Loss	63.24	60.98	49.17	77.85	45.54	58.23	77.36	75.80	74.77	60.85	63.69	62.59	54.74	62.85	53.67	86.04	47.72

Table 1: In-domain results reported with macro averaged F1, averaged over experiments. In lines under *TESTED*, we replace (for Sampling) (→) or remove (for loss) (−), the comprising components.

	F_1 avg.	arc	iac1	perspectrum	poldeb	scd	emergent	fnc1	snopes	mtsd	rumor	semeval16	semeval19	wtwt	argmin	ibmcs	vast
MoLE w/ Hard Mapping	32.78	25.29	35.15	29.55	22.80	16.13	58.49	47.05	29.28	23.34	32.93	37.01	21.85	16.10	34.16	72.93	22.89
MoLE w/ Weak Mapping	49.20	**51.81**	38.97	58.48	47.23	53.96	**82.07**	51.57	56.97	40.13	**51.29**	36.31	31.75	22.75	50.71	75.69	37.15
MoLE w/Soft Mapping	46.56	48.31	32.21	62.73	54.19	51.97	46.86	57.31	53.58	37.88	44.46	36.77	28.92	28.97	57.78	72.11	30.96
TESTED	**59.41**	50.80	**57.95**	**78.95**	**55.62**	**55.23**	80.80	**72.51**	**61.70**	**55.49**	39.44	**40.54**	**46.28**	**42.77**	**72.07**	**86.19**	**54.33**
Topic Sampling → Stratified	50.38	38.47	46.54	69.75	50.54	51.37	68.25	59.41	51.64	48.24	28.04	29.69	34.97	38.13	63.83	83.20	44.06
- Contrastive Loss	54.63	47.96	50.09	76.51	47.49	51.93	75.22	68.69	56.53	49.47	33.95	37.96	44.10	39.56	63.09	83.59	48.03

Table 2: Out-of-domain results with macro averaged F1. In lines under *TESTED*, we replace (for Sampling) (→) or remove (for loss) (−), the comprising components. Results for MoLE w/Soft Mapping are aggregated across with best per-embedding results present in the study (Hardalov et al., 2021).

and *semeval16*, meaning that learning to differentiate between contrastive pairs is essential within this task. We analyse the effect of the contrastive training objective further in subsection 6.4.

Out-of-domain In the out-of-domain evaluation, we leave one dataset out of the training process for subsequent testing. We present the results of TESTED in Table 2, showing that it is able to overperform over the previous state-of-the-art significantly. The metrics in each column of Table 2 show the results for each dataset held out from training and only evaluated on. Our method records an increased performance on 13 of 16 datasets, with an averaged increase of 10.2 F1 points over MoLE, which is a significantly more pronounced increase than for the in-domain setting, demonstrating that the strength of TESTED lies in better out-of-domain generalisation. We can also confirm that replacing the sampling technique or removing the contrastive loss results in lower performance across all datasets, with decreases of 9 and 5 F1 points respectively. This effect is even more pronounced compared to the in-domain experiments, as adapting to unseen domains and topics is facilitated by diverse samples with a balanced label distribution.

6.2 Imbalance Mitigation Through Sampling

Inter-Topic To investigate the inter-topic imbalances, we look at the topic distribution for the top 20 most frequent topics covered in the complete multi-domain dataset $\mathcal{D}$, which accounts for $\geq 40\%$ of the overall data. As we can see in Figure 2, even the most frequent topics greatly vary in their representation frequency, with $\sigma = 4093.55$, where σ is the standard deviation between represented amounts. For the training dataset $\mathcal{D}_{\textbf{train}}$, by contrast, the standard deviation between the topics is much smaller $\sigma = 63.59$. This can be attributed to the fact that $\mathcal{D}_{\textbf{train}}$ constitutes $\leq 10\%$ of $\mathcal{D}$, thus we also show the aggregated data distributions in Figure 2. For a more systematic analysis, we employ the two sample Kolmogorov-Smirnov (KS) test (Massey, 1951), to compare topic distributions in $\mathcal{D}$ and $\mathcal{D}_{\textbf{train}}$ for each dataset present in $\mathcal{D}$. The test compares the cumulative distributions (CDF) of the two groups, in terms of their maximum-absolute difference, stat $= \sup_x |F_1(x) - F_2(x)|$.

The results in Table 3 show that the topic distribution within the full and sampled data $\mathcal{D}$, $\mathcal{D}_{\textbf{train}}$, cannot be the same for most of the datasets. The results for the maximum-absolute difference also show that with at least 0.4 difference in CDF, the

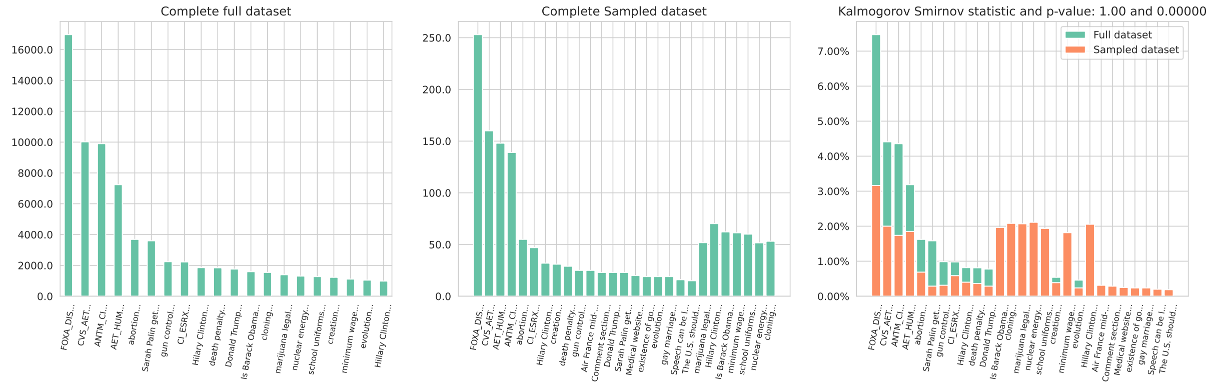

Figure 2: Distributions of top 20 most frequent topics in complete dataset $\mathcal{D}$ (left), Sampled dataset $\mathcal{D}_{\textbf{train}}$ (mid) and their aggregated comparison (right). The distribution of top 20 topics in $\{\mathcal{D}\} - \{\mathcal{D}_{\textbf{train}}\}$ is added to the tail of the figure (mid).

dataset	stat	p-value
fnc-1-ours	1.00	0.007937
arc	0.40	0.873016
emergent	0.80	0.079365
wtwt	0.20	1.000000
rumor	0.40	0.873016
snopes	0.40	0.873016
perspectrum	0.60	0.357143
vast	0.60	0.357143
semeval2016task6	0.40	0.873016
iac	0.40	0.873016
mtsd	0.25	1.000000
argmin	0.40	0.873016
scd	1.00	0.007937
ibm_claim_stance	0.80	0.079365
politicaldebates	0.50	1.000000

Table 3: KS test for topic distributions. The topics in bold designate a rejected null-hypothesis (criteria: $p \leq 0.05$ or $stat \geq 0.4$), that the topics in $\mathcal{D}$ and $\mathcal{D}_{\textbf{train}}$ come from the same distribution.

sampled dataset $\mathcal{D}_{\textbf{train}}$ on average has a more balanced topic distribution. The analysis in Figure 2 and Table 3, show that the sampling technique is able to mitigate the inter-topic imbalances present in $\mathcal{D}$. A more in-depth analysis for each dataset is provided in Appendix A.

Per-topic For the per-topic imbalance analysis, we complete similar steps to the inter-topic analysis, with the difference that we iterate over the top 20 frequent topics looking at *label* imbalances within each topic. We examine the label distribution for the top 20 topics for a per-topic comparison. The standard deviation in label distributions averaged across those 20 topics is $\sigma = 591.05$ for the whole dataset $\mathcal{D}$ and the sampled set $\mathcal{D}_{\textbf{train}}$ $\sigma = 11.7$. This can be attributed to the stratified manner of our sampling technique. This is also

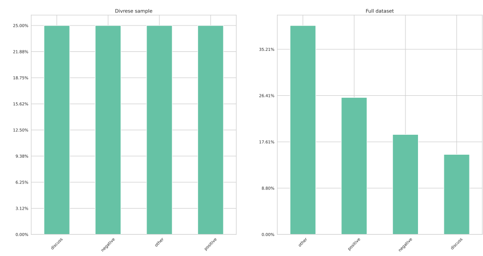

Figure 3: Label distribution in $\mathcal{D}$ (right) and $\mathcal{D}_{\textbf{train}}$ (left).

evident from Figure 3, which portrays the overall label distribution in $\mathcal{D}$ and $\mathcal{D}_{\textbf{train}}$.

To investigate the difference in label distribution for each of the top 20 topics in $\mathcal{D}$, we use the KS test, presented in Table 4. For most topics, we see that the label samples in $\mathcal{D}$ and $\mathcal{D}_{\textbf{train}}$ cannot come from the same distribution. This means that the per-topic label distribution in the sampled dataset $\mathcal{D}_{\textbf{train}}$, does not possess the same imbalances present in $\mathcal{D}$.

We can also see the normalized standard deviation for the label distribution within $\mathcal{D}_{\textbf{train}}$ is lower than in $\mathcal{D}$, as shown in Figure 4. This reinforces the finding that per-topic label distributions in the sampled dataset are more uniform. For complete per-topic results, we refer the reader to Appendix A.

Performance Using our topic-efficient sampling method is highly beneficial for in- and out-of-domain experiments, presented in Table 1 and Table 2. Our sampling method can select diverse and representative examples while outperforming *Random* and *Stratified* sampling techniques by 8 and 5 F1 points on average. This performance can be attributed to the mitigated inter- and per-topic

topic	p-values
FOXA_DIS	0.028571
CVS_AET	0.028571
ANTM_CI	0.028571
AET_HUM	0.047143
abortion	0.100000
Sarah Palin getting divorced?	0.028571
gun control	0.001879
CI_ESRX	0.028571
Hilary Clinton	0.001468
death penalty	0.100000
Donald Trump	0.002494
Is Barack Obama muslim?	0.028571
cloning	0.333333
marijuana legalization	0.032178
nuclear energy	0.333333
school uniforms	0.333333
creation	0.003333
minimum wage	0.333333
evolution	0.100000
lockdowns	0.000491

Table 4: KS test for label distributions. The topics in bold designate a rejected null-hypothesis (criteria: $p \leq 0.05$), that the label samples in $\mathcal{D}$ and $\mathcal{D}_{\textbf{train}}$ averaged per top 20 topics come from the same distribution.

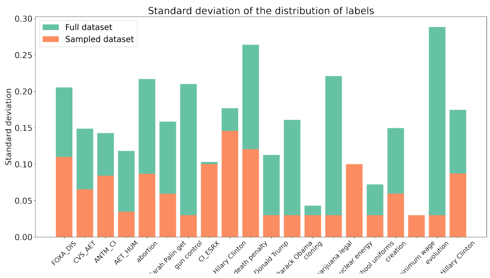

Figure 4: Normalized Standard Deviation in label distribution for top 20 topics.

imbalance in $\mathcal{D}_{\textbf{train}}$.

6.3 Data Efficiency

TESTED allows for sampling topic-efficient, diverse and representative samples while preserving the balance of topics and labels. This enables the training of data-efficient models for stance detection while avoiding redundant or noisy samples. We analyse the data efficiency of our method by training on datasets with sizes $[1\%, 15\%]$ compared to the overall data size $|\mathcal{D}|$, sampled using our technique. Results for the in-domain setting in terms of averaged F1 scores for each sampled dataset size are shown in Figure 5. One can observe a steady performance increase with the more selected samples, but diminishing returns from the 10% point onwards. This leads us to use 10% as the optimal threshold for our sampling process, reinforcing the data-efficient nature of TESTED.

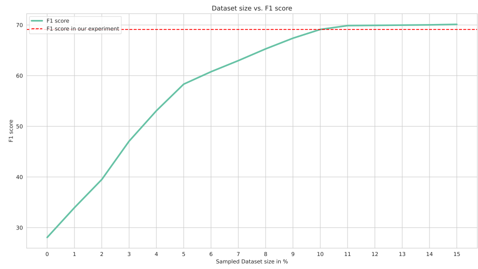

Figure 5: Sampled Data size vs Performance. Performance increases with a bigger sampled selection.

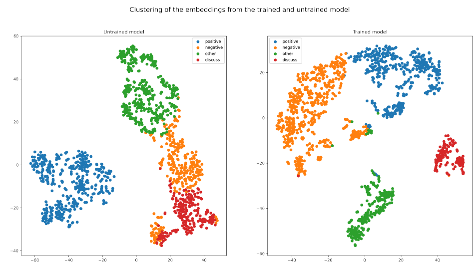

Figure 6: Sample Representation before (left) and after (right) contrastive training.

6.4 Contrastive Objective Analysis

To analyse the effect of the contrastive loss, we sample 200 unseen instances stratified across each dataset and compare the sentence representations before and after training. To compare the representations, we reduce the dimension of the embeddings with t-SNE and cluster them with standard K-means. We see in Figure 6 that using the objective allows for segmenting contrastive examples in a more pronounced way. The cluster purity also massively rises from 0.312 to 0.776 after training with the contrastive loss. This allows the stance detection model to differentiate and reason over the contrastive samples with greater confidence.

7 Conclusions

We proposed TESTED, a novel end-to-end framework for multi-domain stance detection. The method consists of a data-efficient topic-guided sampling module, that mitigates the imbalances inherent in the data while selecting diverse examples, and a stance detection model with a contrastive training objective. TESTED yields significant performance gains compared to strong baselines on in-domain experiments, but in particular generalises well on out-of-domain topics, achieving a 10.2 F1 point improvement over the state of the art, all while using $\leq 10\%$ of the training data. While in this paper, we have evaluated TESTED on stance detection, the method is applicable to text classification more broadly, which we plan to investigate in more depth in future work.

Limitations

Our framework currently only supports English, thus not allowing us to complete a cross-lingual study. Future work should focus on extending this study to a multilingual setup. Our method is evaluated on a 16 dataset stance benchmark, where some domains bear similarities. The benchmark should be extended and analyzed further to find independent datasets with varying domains and minimal similarities, allowing for a more granular out-of-domain evaluation.

Acknowledgements

This research is funded by a DFF Sapere Aude research leader grant under grant agreement No 0171-00034B, as well as supported by the Pioneer Centre for AI, DNRF grant number P1.

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.

Abeer Aldayel and Walid Magdy. 2019. Your Stance is Exposed! Analysing Possible Factors for Stance Detection on Social Media. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *ArXiv preprint*, abs/2008.09470.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Bjorn Barz and Joachim Denzler. 2020. Deep learning on small datasets without pre-training using cosine loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1371–1380.

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. 2009. Importance weighted active learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 49–56. ACM.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2001. Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 601–608. MIT Press.

Filip Boltužić and Jan Šnajder. 2014. Back up your Stance: Recognizing Arguments in Online Discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, Maryland. Association for Computational Linguistics.

Felix Buchert, Nassir Navab, and Seong Tae Kim. 2022. Exploiting Diversity of Unlabeled Data for Label-Efficient Semi-Supervised Active Learning. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2063–2069. IEEE.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. 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 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online. Association for Computational Linguistics.

Alexis Conneau, 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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Bo Dai and Dahua Lin. 2017. Contrastive Learning for Image Captioning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 898–907.

Rajshekhar Das, Yu-Xiong Wang, and José MF Moura. 2021. On the importance of distractors for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9030–9040.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. 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 879–895, Online. Association for Computational Linguistics.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. 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 1596–1611, Online. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants. 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 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: Word-level Adversarial ReProgramming. 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 4921–4933, Online. 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.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022a. A Survey on Stance Detection for Mis- and Disinformation Identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022b. Few-Shot Cross-Lingual Stance Detection with Sentiment-Based Pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10729–10737.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, Nagoya, Japan. Asian Federation of Natural Language Processing.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are You Taking this Stance? Identifying and Classifying Reasons in Ideological Debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar. Association for Computational Linguistics.

Hideitsu Hino. 2020. Active learning: Problem settings and recent developments. *ArXiv preprint*, abs/2012.04225.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. 2017. Learning Active Learning from Data. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4225–4235.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv preprint*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Frank J. Massey. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253):68–78.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Moritz Osnabrügge, Elliott Ash, and Massimo Morelli. 2023. Cross-domain topic classification for political texts. *Political Analysis*, 31(1):59–80.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Advances in Neural Information Processing Systems*, 34:11054–11070.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge stage 1 (FNC-I): Stance detection. *URL www. fakenewschallenge. org*.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying Misinformation in Microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Tacoma Tacoma, Hao Li, and Rong Jin. 2019. Softtriple loss: Deep metric learning without triplet sampling. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6449–6457. IEEE.

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

Hiranmayi Ranganathan, Hemanth Venkateswara, Shayok Chakraborty, and Sethuraman Panchanathan. 2017. Deep active learning for image classification.

In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3934–3938. IEEE.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40.

Nils Rethmeier and Isabelle Augenstein. 2023. A Primer on Contrastive Pretraining in Language Processing: Methods, Lessons Learned, and Perspectives. *ACM Comput. Surv.*, 55(10).

Myrthe Reuver, Antske Fokkens, and Suzan Verberne. 2021. No NLP task should be an island: Multi-disciplinarity for diversity in news recommender systems. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 45–55, Online. Association for Computational Linguistics.

Jingyu Shao, Qing Wang, and Fangbing Liu. 2019. Learning to sample: an active learning framework. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 538–547. IEEE.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From Argumentation Mining to Stance Classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77, Denver, CO. Association for Computational Linguistics.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing Stances in Ideological On-Line Debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, Los Angeles, CA. 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.

Ali Raza Syed, Andrew Rosenberg, and Ellen Kislal. 2016. Supervised and unsupervised active learning for automatic speech recognition of low-resource languages. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pages 5320–5324. IEEE.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A Corpus for Research on Deliberation and Debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817, Istanbul, Turkey. European Language Resources Association (ELRA).

Rui Wang, Deyu Zhou, Mingmin Jiang, Jiasheng Si, and Yang Yang. 2019. A survey on opinion mining: From stance to product aspect. *IEEE Access*, 7:41101–41124.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Penghui Wei, Junjie Lin, and Wenji Mao. 2018. Multi-Target Stance Detection via a Dynamic Memory-Augmented Network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1229–1232. ACM.

Penghui Wei and Wenji Mao. 2019. Modeling Transferable Topics for Cross-Target Stance Detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1173–1176. ACM.

Xuyang Yan, Shabnam Nazmi, Biniam Gebru, Mohd Anwar, Abdollah Homaifar, Mrinmoy Sarkar, and Kishor Datta Gupta. 2022. Mitigating shortage of labeled data using clustering-based active learning with diversity exploration. *ArXiv preprint*, abs/2207.02964.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. 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 5065–5075, Online. Association for Computational Linguistics.

Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. 2015. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127.

Yi Yang, Shimei Pan, Doug Downey, and Kunpeng Zhang. 2014. Active learning with constrained topic model. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 30–33, Baltimore, Maryland, USA. Association for Computational Linguistics.

Giulio Zhou and Gerasimos Lampouras. 2021. Informed sampling for diversity in concept-to-text NLG. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2494–2509, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1137–1144, Manchester, UK. Coling 2008 Organizing Committee.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018a. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018b. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing and Management*, 54(2):273–290.

Appendix

A Imbalance analysis

A.1 Inter-topic

To complement our inter-topic imbalance mitigation study, we complete an ablation on all topics in $\mathcal{D}$ and report them on a per-domain basis in Figure 7. The trend is similar to the one in Figure 2, where the dataset with imbalanced distributions is rebalanced, and balanced datasets are not corrupted.

A.2 Per-topic

We show that our topic-efficient sampling method allows us to balance the label distribution for unbalanced topics, while not corrupting the ones distributed almost uniformly. To do this, we investigate each of the per-topic label distributions for the top 20 most frequent topics while comparing the label distributions for $\mathcal{D}$ and $\mathcal{D}_{\textbf{train}}$, presented in Figure 8.

B Evaluation Metrics

To evaluate our models and have a fair comparison with the introduced benchmarks we use a standard set of metrics for classification tasks such as macro-averaged F1, precision, recall and accuracy.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Prec = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = \frac{2 * Prec * Recall}{Prec + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{7}$$

C Dataset Statistics

We use a stance detection benchmark (Hardalov et al., 2021) whose data statistics are shown in Table 5. The label mapping employed is shown in Table 6.

D TESTED with different backbones

We chose to employ different PLM's as the backbone for TESTED and report the results in the Table 7. The PLMs are taken from the set of *roberta-base, roberta-large, xlm-roberta-base, xlm-roberta-large*. The differences between models with a similar number of parameters are marginal. We can see a degradation of the F1 score between the *base* and *large* versions of the models, which can be attributed to the expressiveness the models possess. We also experiment with the distilled version of the model and can confirm that in terms of the final F1 score, it works on par with the larger models. This shows that we can utilise smaller and more computationally efficient models within the task with marginal degradation in overall performance.

Dataset	Train	Dev	Test	Total
arc	12,382	1,851	3,559	17,792
argmin	6,845	1,568	2,726	11,139
emergent	1,770	301	524	2,595
fnc1	42,476	7,496	25,413	75,385
iac1	4,227	454	924	5,605
ibmcs	935	104	1,355	2,394
mtsd	3,718	520	1,092	$5,330$
perspectrum	6,978	2,071	2,773	11,822
poldeb	4,753	1,151	1,230	7,134
rumor	6,093	471	505	$7,276$
scd	3,251	624	964	4,839
semeval2016t6	2,497	417	1,249	4,163
semeval2019t7	5,217	1,485	1,827	8,529
snopes	14,416	1,868	3,154	19,438
vast	13,477	2,062	3,006	18,545
wtwt	25,193	7,897	18,194	51,284
Total	154,228	30,547	68,495	253,270

Table 5: Dataset statistics of the stance detection benchmark by Hardalov et al. (2021) also used in this paper. Note that the rumour and mtsd datasets are altered in that benchmark as some of the data was unavailable.

Label	Description
Positive	agree, argument for, for, pro, favor, support, endorse
Negative	disagree, argument against, against, anti, con, undermine, deny, refute
Discuss	discuss, observing, question, query, comment
Other	unrelated, none, comment
Neutral	neutral

Table 6: Hard stance label mapping employed in this paper, following the stance detection benchmark by Hardalov et al. (2021).

Figure 7: Distributions of top 20 most frequent topics for each dataset (left), Sampled dataset $\mathcal{D}_{\textbf{train}=dataset}$ (mid) and their aggregated comparison (right).

Figure 8: Distributions of labels for top 20 most frequent topics for $\mathcal{D}$ (left), Sampled dataset $\mathcal{D}_{\textbf{train}=dataset}$ (mid) and their aggregated comparison (right).

	F_1 avg.	arc	iac1	perspectrum	poldeb	scd	emergent	fnc1	snopes	mtsd	rumor	semeval16	semeval19	wtwt	argmin	ibmcs	vast
TESTED$_{reberta\text{-}large}$	69.12	64.82	56.97	83.11	52.76	64.71	82.10	83.17	78.61	63.96	66.58	69.91	58.72	70.98	62.79	88.06	57.47
TESTED$_{xlm\text{-}reberta\text{-}large}$	68.86	64.35	57.0	82.71	52.93	64.75	81.72	82.71	78.38	63.66	66.71	69.76	58.27	71.29	62.73	87.75	57.2
TESTED$_{reberta\text{-}base}$	65.32	59.71	51.86	76.75	50.23	61.35	78.84	82.09	73.31	62.87	65.46	63.89	58.3	67.28	58.28	83.81	51.09
TESTED$_{xlm\text{-}reberta\text{-}base}$	65.05	60.26	51.96	76.2	51.82	58.74	74.68	77.9	72.61	62.71	66.08	69.74	53.27	65.83	59.09	87.92	52.08
TESTED$_{distilroberta\text{-}base}$	68.86	61.78	56.94	80.36	46.29	64.1	79.26	81.37	73.44	62.6	63.4	63.75	56.53	68.35	57.27	81.93	56.3

Table 7: In-domain results reported with macro averaged F1, with varying backbones when using TESTED.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix D

C ☑ Did you run computational experiments?

6

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We use standard pre-trained language models.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
6

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.