Transfer and Active Learning for Dissonance Detection: Addressing the Rare-Class Challenge

Vasudha Varadarajan[♣*], **Swanie Juhng**[♣*], **Syeda Mahwish**[♣], **Xiaoran Liu**[♠]
Jonah Luby, Christian C. Luhmann[♠] and **H. Andrew Schwartz**[♣]
[♣]Department of Computer Science, [♠]Department of Psychology
Stony Brook University
{vvaradarajan,sjuhng,smahwish,has}@cs.stonybrook.edu
{christian.luhmann,xiaoran.liu}@stonybrook.edu, jonahluby@gmail.com

Abstract

While transformer-based systems have enabled greater accuracies with fewer training examples, data acquisition obstacles still persist for rare-class tasks – when the class label is very infrequent (e.g., < 5% of samples). Active learning has in general been proposed to alleviate such challenges, but choice of selection strategy, the criteria by which rare-class examples are chosen, has not been systematically evaluated. Further, transformers enable iterative transfer-learning approaches. We propose and investigate transfer- and active learning solutions to the rare class problem of dissonance detection through utilizing models trained on closely related tasks and the evaluation of acquisition strategies, including a proposed *probability-of-rare-class* (PRC) approach. We perform these experiments for a specific rare class problem: collecting language samples of cognitive dissonance from social media. We find that PRC is a simple and effective strategy to guide annotations and ultimately improve model accuracy, and while transfer-learning in a specific order can improve the cold-start performance of the learner but does not benefit iterations of active learning.

1 Introduction

Cognitive dissonance occurs during everyday thinking when one experiences two or more beliefs that are inconsistent in some way (Harmon-Jones and Harmon-Jones, 2007). Often expressed in language, dissonance plays a role in many aspects of life, for example affecting health-related behavior such as smoking (Chapman et al., 1993) and contributing to the development of (and exit from) extremism (Dalgaard-Nielsen, 2013). However, while the phenomenon is common enough to occur on a daily basis, dissonance is still relatively rare among the myriad of other relationships between

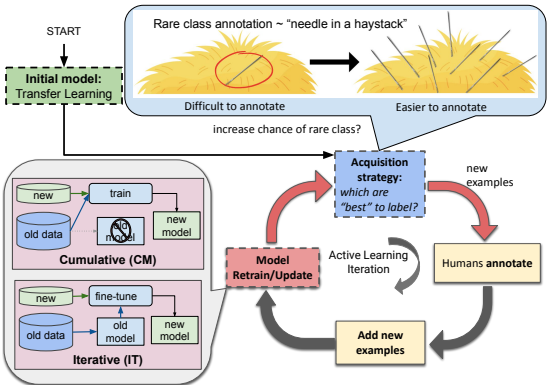

Figure 1: Demonstration of the active learning (AL) loop in general. Our paper examines the three highlighted steps: (i) Bootstrapping with TL model, (ii) Acquisition strategy, and (iii) Model update.

beliefs that occur across random selections of linguistic expressions and thus makes the automatic detection of it a rare-class problem.

Despite recent advances in modeling sequences of words, rare-class tasks – when the class label is very infrequent (e.g., < 5% of samples) – remain challenging due to the low rate of positive examples. Not only are more random examples necessary to reach a substantial amount of the rare class (e.g., 1,000 examples to reach just 50 examples), but also it is easy for human annotators to miss the rare instances where dissonance is present. Here, we develop and address the challenges of creating a resource for language-based assessment of dissonance.

Active learning using large language models presents both new opportunities and challenges. On the one hand, large language models (LLMs) offer unmatched representations of documentations, able to achieve state-of-the-art language understanding task performance with transfer learning, often only with a few iterations of fine-tuning (Liu et al., 2019). On the other hand, representations are high-dimensional, and models trained or fine-tuned with only a small number of examples are prone to over-

* co-lead authors

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 11923–11936
July 9-14, 2023

fitting, especially when there is a large class imbalance as in rare-class problems. While LLMs have enabled attempts to tackle increasingly complex semantic challenges across a growing list of tasks, getting annotated examples for such problems can become a bottleneck due to its time- and labor-intensiveness (Wu et al., 2022). Since data-centric improvements for more novel tasks can provide a faster path than model-centric improvements (Ng, 2021), active learning can be a way forward to be both data-centric and address bottlenecks in label acquisition – it aims to reduce annotation costs as well as alleviate the training data deficiency that large language models face.

However, while active learning has been studied for multiple natural language tasks (Shen et al., 2017; Liang et al., 2019), little is known about active learning acquisition strategies for LM-based approaches, especially for rare-class problems. *High data imbalance* coupled with *very less training data* poses the challenge of "absolute rarity" (Al-Stouhi and Reddy, 2016), as in our task of dissonance detection. We address this problem by using a novel combination of evaluating the ordering of transfer learning from similar tasks to cold-start the active learning loop, and by acquiring with a relatively simple acquisition strategy focused on *probability-of-rare-class* (PRC) to increase the rare class samples.

Our contributions include: (1) finding that bootstrapping AL models with transfer learning on closely related tasks significantly improves rare class detection; (2) a novel systematic comparison of five common acquisition strategies for active learning for a rare class problem[1]; (3) a systematic comparison of two different approaches to handling AL iterations for LLMs – cumulative and iterative fine-tuned model updates – finding the cumulative approach works best; (4) evaluating annotation costs of a rare-class task, finding that minimum annotation cost does not necessarily lead to better models, especially in realistic scenarios such as *absolute rarity*; and (5) release of a novel dataset[2] for the task of identifying cognitive dissonance in social media documents.

[1]Code: https://github.com/humanlab/rare-class-AL

[2]Dataset: https://github.com/humanlab/dissonance-twitter-dataset

2 Related Work

Active learning in NLP has been largely studied as a theoretical improvement over traditional ML for scarce data. In this work, we specifically investigate *pool-based* active learning, or picking out samples to annotate from a larger pool of unlabeled data, and particularly data for a *rare-class* problem where LMs are not well-understood yet.

Acquisition strategies Sampling strategies for active learning can be broadly classified into three: uncertainty sampling (Shannon, 1948; Wang and Shang, 2014; Netzer et al., 2011), representative (or diversity) sampling (Citovsky et al., 2021; Sener and Savarese, 2018; Gissin and Shalev-Shwartz, 2019), and the combination of the two (Zhan et al., 2022). The uncertainty sampling strategies that employ classification probabilities, Bayesian methods such as variational ratios (Freeman, 1965), and deep-learning specific methods (Houlsby et al., 2011) often use epistemic (or model) uncertainty. We choose maximum entropy to represent the uncertainty sampling, since it is usually on par with more elaborated counterparts (Tsvigun et al., 2022). As a popular diversity sampling baseline to compare against, we pick select CoreSet (Sener and Savarese, 2018). The state-of-the-art methods combine these two strategies in novel ways, such as using statistical uncertainty in combination with some form of data clustering for diversity sampling (Zhang and Plank, 2021; Ash et al., 2019). Our work uses Contrastive Active Learning (Margatina et al., 2021) to represent this strategy.

On the other hand, Karamcheti et al. (2021) and Munjal et al. (2022) claim there is rather small to no advantage in using active learning strategies, because a number of samples might be collectively outliers, and existing strategies contribute little to discover them and instead harm the performance of subsequent models. Researchers recently have also focused on the futility of complex acquisition functions applied to difficult problems and argued that random acquisition performs competitive to more sophisticated strategies, especially when the labeled pool has grown larger (Sener and Savarese, 2018; Ducoffe and Precioso, 2018). Furthermore, a large-scale annotation of randomly sampled data could be less expensive than ranking data to annotate in each round of active learning, if there is not much advantage (i.e., such as capturing rare classes) in using a specific strategy.

Cold-Start AL While the problem of cold-start exists in acquiring samples through active learning, some work has been done to combat this by leveraging the learned weights in pretrained models (Yuan et al., 2020). However, there is much to gain from the field of transfer learning especially for rare class problems, as seen in Al-Stouhi and Reddy (2016). We borrow the concept of heterogeneous transfer learning (Day and Khoshgoftaar, 2017; Zhuang et al., 2021) and transfer the model weights directly obtained from pretraining on closely related (but different) tasks on completely different domains. This helps models to improve the zero-shot ability for rare class detection. Such methods have been explored in traditional machine learning (Kale and Liu, 2013) but not in the era of large language models to the best of our knowledge.

Rare class AL There has been a growing number of applications of active learning in data imbalance and rare class problems. Such works include (Kothawade et al., 2021; Choi et al., 2021; Ein-Dor et al., 2020) which proposed frameworks to improve model performance with data imbalance but failed to check the feasibility and costs in a real-world, active annotation setting where not only is rare class very infrequent (4%) but very few (< 70) examples of the rare class exist due to small dataset size ("absolute rarity"). They also fail to compare against a simple, rare class probability of the model. While some work in the pre-LLM era use probability outputs of a classifier (certainty-based sampling) which is similar to the proposed PRC, they claim to work better in conjunction with co-selection using other uncertainty sampling strategies, and that certainty-based sampling alone performs poorly in terms of increasing rare-class samples (Li et al., 2012). Many studies also focus on rare class *discovery*, or finding outlying samples that do not fall under the existing categories (Hospedales et al., 2013; Haines and Xiang, 2014; Hartford et al., 2020). This is different from our task which focuses on the *detection* of a rare class.

3 Task

Cognitive dissonance is a phenomenon that happens when two elements of cognition (i.e., thoughts, experiences, actions, beliefs) within a person do not follow one another or are contradictory, and consonance is when one belief follows from the other (Harmon-Jones and Mills, 2019). Cognitive dissonance raises psychological discomfort, encouraging a person to resolve the dissonance. As the magnitude of dissonance increases, the pressure to resolve it grows as well (Harmon-Jones et al., 2008; McGrath, 2017).

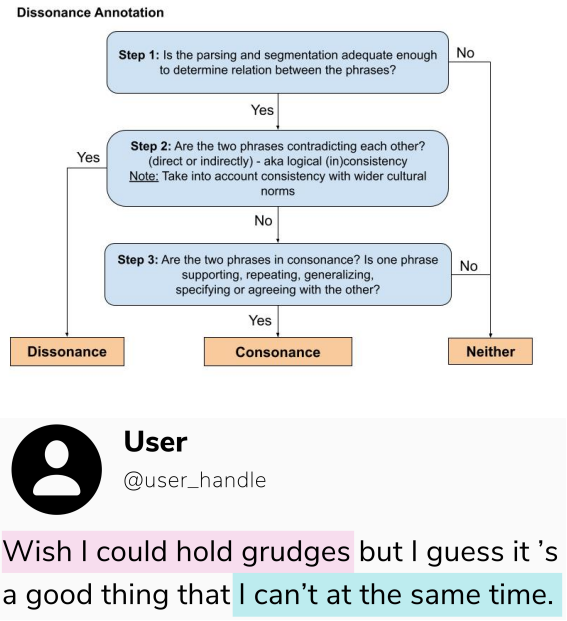

Figure 2: Above: Flowchart describing the steps for the annotators to label tweets as DISSONANCE, CONSONANCE, or NEITHER.
Below: An example of a pair of THOUGHT segments in a tweet annotated as dissonance.

Social psychology has used this human tendency to resolve dissonance to understand important psychological processes such as determinants of attitudes and beliefs, consequences of decisions, internalization of values, and the effects of disagreement among persons (Harmon-Jones and Mills, 2019). Dissonance is also related to anxiety disorders (Juhng et al., 2023), relevant to understanding extremism and predicting cognitive styles of users. Our approach to annotating cognitive dissonance on social media is motivated by the two-stage annotation approach described in (Varadarajan et al., 2022). To the best of our knowledge this is the first social media dataset for cognitive dissonance.

4 Methods

4.1 Annotation and Dataset

Following the definition of cognitive dissonance in §3, we treat discourse units as semantic elements that can represent beliefs. A discourse unit consists of words or phrases that have a meaning (Polanyi, 1988)– and then cognitive dissonance is analogous

to a discourse relation between two discourse units. Recent work (Son et al., 2022) represents discourse relations in a continuous vector space, motivating us to look at cognitive dissonance, too, as a relationship between two "thought" discourse units.

We build a dissonance dataset by first sampling posts between 2011 and 2020 on Twitter. The tweets were parsed into discourse units using the parser by Wang et al. (2018) which uses the PDTB framework. [3]

Each discourse unit in a document is initially annotated into THOUGHT or OTHER. [4] A THOUGHT is a discourse unit describing the author's own beliefs, experiences and actions and are potential elements to be in dissonance. OTHER comprises of anything else, from meaningless phrases to coherent beliefs that belong to someone other than the author. For the annotation of dissonance, pairs of THOUGHT units from each tweet are extracted, and then annotated to compose CONSONANCE, DISSONANCE or NEITHER according to the framework described in Figure 2 – a three-class annotation. This framework was developed from annotator training to spot examples of dissonance, followed by discussion with a cognitive scientist.

Among a random selection of tweets, the natural frequency of the DISSONANCE class is around 3.5%. The annotations were carried out by a team of three annotators, with the third annotator tiebreaking the samples disagreed by the first two annotators.

Initial set ($iter_0$) This dataset is used to select the best transfer model to effectively cold-start the AL loop. We start with a total of 1,901 examples of dissonance task annotations, which we split into a training set of 901 examples (henceforth, $iter_0$) with 43 examples of dissonance (4.77%) picked randomly from discourse-parsed tweets. We create initial development and test sets with 500 examples each. They were created such that all the THOUGHT pairs that were a part of a single tweet belong to the same set.

Final development and test datasets We gather additional 984 annotations for development set and 956 annotations for test set in addition to the previously mentioned 500 for each, summing up to 1,484 development examples (dev) and 1,456 test examples ($test$) with around 10% dissonance examples in each, to account for increased frequency of occurrence of the rare class after incorporating novel acquisition strategies.

4.2 Modeling

4.2.1 Architecture

A RoBERTa-based dissonance classifier is used consistently across all the experiments in this paper: for any two THOUGHT segments belonging to a single post, the input is in the form of "[CLS] $segment_1$ [SEP] $segment_2$ [SEP]". We take the contextualized word embedding $\mathbf{x} \in \mathbb{R}^d$ of [CLS] in the final layer and feed it into the linear classifier: $y = \text{softmax}(W\mathbf{x} + \mathbf{b})$, where $W \in \mathbb{R}^{d\times 2}, \mathbf{b} \in \mathbb{R}^2$ is a learned parameter. We trained the model parameters with cross entropy loss for 10 epochs, using AdamW optimizer with the learning rate of 3×10^{-5}, batch size of 16, and warm up ratio of 0.1. To avoid overfitting, we use early stopping (patience of 4) with the AUC score. We run the AL experiments on the datasets delineated in §4.1.

While the annotations are for three classes (Figure 2), the models used for AL across all strategies classify labels to binary level (dissonance or not dissonance), as we are focused specifically on the dissonance class – while dissonance is rare, it is also essential to perform well in detecting this class.

4.2.2 Bootstrap with Transfer Learning

We explore cold-starting the active annotation process using a transfer of model weights trained on similar tasks.

PDTB-Comparison/Expansion (CE) The PDTB framework defines discourse relations at three hierarchies: Classes, Types and Subtypes. Of the four classes viz. Temporal, Contingency, Comparison, and Expansion, the Comparison class "indicates that a discourse relation is established between two discourse units in order to highlight prominent differences between the two situations" (Prasad et al., 2008). While this class is different from DISSONANCE, it is useful in capturing

[3] PDTB (Prasad et al., 2008) and RST (Mann and Thompson, 1987) are the two major frameworks for discourse parsing; we use the former for this work since PDTB is lexically grounded and identifies discourse relations using lexical cues. While the RST framework could be helpful since rhetorical relations are viewed as cognitive entities (Taboada and Mann, 2006), the complex relationships defined with RST's nested structures can complicate our search for cognitive dissonance samples at the preliminary stage of data collection.

[4] This was a simpler, large-scale annotation to pick out discourse units describing author's own beliefs. We do not go into the details of this specific annotation since it is not pertinent to this work.

discord between the semantics of two discourse units. The Expansion class is defined to "cover those relations which expand the discourse and move its narrative or exposition forward," which is closer to our conception of CONSONANCE. We thus identify a similar task to be classifying discourse relations as Comparison or Expansion (CE). The CE dataset consists of 8,394 (35.12%) in Comparison class and 15,506 (64.88%) in Expansion class. The model was trained on the architecture as explained in §4.2.1 with $segment_1$ as the first discourse unit (Arg1) and $segment_2$ as the second discourse unit (Arg2) and the output indicating Comparison or Expansion class. For the training, 10% was set aside as the development set to pick the best performing model on the CE task.

	F1-macro	F1-Dis	Prec-Dis	Rec-Dis	AUC
Diss alone	0.478	0.000	0.000	0.000	0.500
Debate	**0.595**	**0.319**	0.349	**0.278**	**0.620**
CE	0.487	0.210	**0.558**	0.129	0.602
Deb; CE	0.540	0.211	0.349	0.152	0.583

Table 1: Performance of models pretrained on two similar tasks, separately and combined, based on development and test set from $iter_0$. Precision and Recall reported for Dissonance class. ";" refers to combining the two datasets. **Bold** represents best in column. Training with dissonance dataset alone doesn't help the model– this shows the usefulness of transfer learning to cold-start active learning, especially on transfer from Debate.

Dissonant Stance Detection (Debate) The dissonant stance detection task classifies two statements in a debate to be in agreement (consonant stance) or disagreement (dissonant stance) independent of the topic that is being debated upon as described in Varadarajan et al. (2022). Dissonant stance is different from DISSONANCE in two ways: (a) each input segment is a complete post consisting of multiple sentences arguing for a stance/topic whereas in our task, they are discourse units; and (b) while both are social media domains, our task uses a more personal, informal language while debate forums use impersonal language citing facts, not author's subjective beliefs. But the tasks are similar in the detection of dissonance between two segments, and we identify it as a potential task to transfer learn from. The statements were extracted from a debate forum consisting of 34 topics with 700 examples each (total 23,800 samples). There were 8,289 dissonant stance examples (34.82%) in the dataset. While the dataset has three labels – consonant stance, dissonant stance and neither –, we train a binary classifier on top of the RoBERTa layers to detect dissonant stance or not dissonant stance, keeping the task similar to the model we use in the AL iterations.

	F1-macro	F1-Dis	Prec-Dis	Rec-Dis	AUC
Transfer-Learning Alone					
Deb; CE	0.520	0.212	0.442	0.140	0.593
Deb→CE	0.495	0.170	0.349	0.112	0.544
CE→Deb	0.487	0.243	0.744	0.146	**0.666**
Transfer and Continue Training					
Deb;CE;$iter_0$	0.458	0.033	0.100	0.020	0.507
Deb→$iter_0$	0.564	0.296	0.236	0.400	0.554
Deb→CE→$iter_0$	0.532	0.143	0.146	0.140	0.531
CE→Deb→$iter_0$	0.585	0.229	0.296	0.186	**0.572**

Table 2: The zero-shot performance of models further fine-tuned from those in Table 1. ";" refers to combining the two datasets, "→" indicates iteratively fine-tuning on each task, and **bold** represents the best in column. Scores based on development and test set from $iter_0$. The order matters for fine-tuning: we find the CE→Deb performs the best in zero-shot setting.

Both of these tasks involve two statements/phrases as inputs, and the output is Comparison/Expansion in the first case, and Dissonant/Not Dissonant stance in the second case. We transfer all the weights of the RoBERTa-base model, leaving out the binary classifier layer when fine-tuning to the cognitive dissonance task. The results of fine-tuning on one or both best transfer model was picked as the model having trained on PDTB and then further fine-tuned on the Debate task as well, as shown in Table 1.

4.2.3 AL strategies

Since our annotation process brought about only a small incremental improvement for performance on the rare class, yet contributed much to modeling the dominant classes, we hypothesized that using probability of the rare class as an acquisition strategy in active learning could work just as well as other strategies that are based on diversity and uncertainty sampling. We ran our analyses over four other common acquisition strategies by picking out the top 10% (300 out of an unannotated data pool containing 3,000 examples). We limit to only four other strategies because of the annotation costs and limited time.

PRC For a rare, hard class, we use a binary classifier that outputs the probability of rare class learned from the samples encountered so far. This is a computationally inexpensive and simple method that could be easily surpassed by other complex AL strategies but was surprisingly found to be the most effective in this study. The examples from the data pool that are predicted to have the highest probability of the rare class by the classification model from previous iteration are selected.

RANDOM As a baseline, we randomly sample examples from the data pool, which reflects the natural distribution of classes. Random method has been considered to be a solid baseline to compare against, as many AL strategies do not merit when the annotation pool scales up and collective outliers are missed, as explained in §2.

ENTROPY We use predictive entropy as the uncertainty-based sampling baseline to compare against. While Least Confident Class (LCC) is a popular strategy to capture samples based on uncertainty, it is calculated based on only one class, working best for binary classification and provides merit within balanced classes, whereas predictive entropy is a generalized form of LCC, and a more popular variant (Freeman, 1965).

CAL Contrastive Active Learning (Margatina et al., 2021) is a state-of-the-art approach that chooses data points that are closely located in the model feature space yet predicted by models to have maximally different likelihoods from each other. This method is relevant to the task at hand because in rare class problems, it is often difficult for a model to learn the decision boundary around the rare class due to the low number of such samples. Thus we focus on a method that tries to pick out samples at the decision boundary of the rare class.

CORESET An acquisition method that has worked well as a diversity sampling method is CoreSet (Sener and Savarese, 2018). This method uses a greedy strategy to sample a subset of data that is most representative of the real dataset, i.e., the larger data pool that we sample from.

4.2.4 Model Update

To the best of our knowledge, the question of model update in an AL loop has not been explored. We explore two fine-tuning approaches to update the model following annotation of new samples in each round of the active learning loop – cumulative (CM) and iterative (IT). Figure 1 provides a visual explanation of the two approaches.

	Random	Entropy	CoreSet	CAL	PRC
Random	×	12.15%	11.52%	10.83%	11.02%
Entropy		×	64.68%	76.33%	87.98%
CoreSet			×	58.67%	61.65%
CAL				×	82.98%

Table 3: % overlap in the samples picked out by the base model for the five strategies described in 4.2.3. Probability of rare class (PRC) has a significant overlap with a state-of-the-art approaches, implying that for the rare class problem, PRC is a computationally inexpensive, alternative acquisition approach.

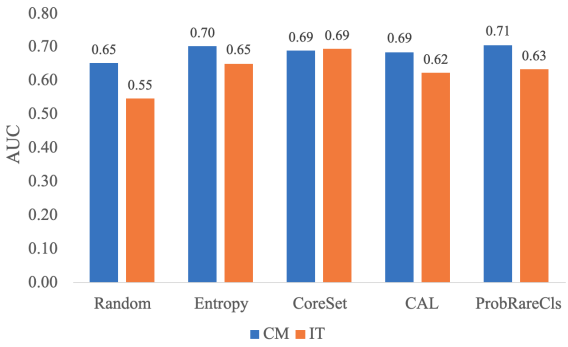

Figure 3: AUC for the five strategies for IT and CM model updates. This shows that the CM model update always performs equally with or better than the IT update.

Cumulative (CM) At each round of the AL loop, the 300 newly annotated samples are combined with the previous ones as the input to fine-tune the classification model from a base pretrained language model.

Iterative (IT) At each round of the AL loop, the 300 newly annotated samples are used to further fine-tune the model trained during the previous loop.

5 Results

5.1 Transfer Learning Models for Cognitive Dissonance

Table 1 displays the evaluation of the transfer learning models on bootstrapping the active annotation, revealing that pretraining the large language models on relevant tasks that are specifically designed to mimic the task at hand can lead to better performance. In addition, the transfer from both Debate

	IT					CM				
Strategy	F1-macro	F1-Dis	Prec-Dis	Rec-Dis	AUC	F1-macro	F1-Dis	Prec-Dis	Rec-Dis	AUC
RANDOM	0.556	0.175	0.119	**0.336**	0.546	0.640	0.362	0.397	0.334	0.652
ENTROPY	0.632	0.351	0.401	0.318	0.650	0.649	**0.398**	0.540	0.315	0.702
CORESET	**0.652**	0.397	0.513	0.329	0.694	0.635	0.375	0.523	0.292	0.688
CAL	0.612	0.306	0.331	0.321	0.623	0.644	0.383	0.497	0.313	0.685
PRC	0.616	0.322	0.371	0.309	0.633	0.633	0.382	**0.580**	0.285	**0.706**

Table 4: Comparison of five annotation strategies for iterative (IT) and cumulative (CM) approaches for 2 class classification. The metrics are averaged over two iterations of active learning, with 300 new examples annotated in each iteration (adds between 3-10% samples of dissonance in each round, depending on the strategy). **Bold** represents the best for each reported metric. The performance of CM approach exceeds that of IT across most acquisition functions, which contrasts with the case of transfer learning step where combining the datasets into one did not help the model as much. While the performance on adding 10 to 30 samples of dissonance is not expected to cause large jumps in performance, note that using the PRC strategy leads to significant gain in performance in detecting the dissonance class compared to the transfer models from Table 2.

and CE tasks leads to better results than training the RoBERTa-base model on the Dissonance dataset directly. We also combine the two datasets used in Debate and CE and train them at the same time – similar to the CM approach – to find that Deb;CE model still performs better than the model directly trained on the Dissonance dataset. This shows the incredible zero-shot abilities of transfer models for this task.

Furthermore, we explore if continuing to pretrain on a different task after already having pretrained on Debate or CE makes a difference. In such case, order of pretraining tasks matters, and there is a much larger gain in the zero-shot performance for CE→Debate compared to Debate→CE as seen in Table 2. When any of these transfer models is further fine-tuned on the dissonance dataset, we find an initial drop in performance. This is explained with the effect of the heterogeneous domain transfer and the small dataset in the $iter_0$ train set. As later shown in Table 4, the performance improves when more samples are collected in the AL iterations. The domain transfer from both tasks (or a combination of them) gives the active annotation a head-start for initial sample selection.

5.2 Acquisition Strategies

Table 3 shows the overlap of samples picked out in each iteration from the same larger data pool for the model at iteration 0 (base model). RANDOM has the lowest overlaps with all the other strategies. We also find that there is a significant overlap ($> 80\%$) in the samples between ENTROPY or CAL, the state-of-the-art approach, and PRC. CAL has a higher overlap with ENTROPY rather than CORESET, showing that samples deemed to be both highly informative and contrastive by the model are also usually likely to be dissonant. This is contrastive to the prior literature revealing that poor calibration of large language models often renders the models to rarely be uncertain of their outcomes (Guo et al., 2017). All strategies except RANDOM have $> 55\%$ overlap with each other. This implies that diversity- and uncertainty-based methods are not as different from each other as they theoretically are and inclined to pick similar samples – hinting that a lot of diversity-based sampling measures mostly pick highly informative samples as well. Furthermore, PRC tends to choose samples that the "state-of-the-art" model also picks in rare-case scenarios, indicating that it could be a computationally inexpensive alternative.

Table 4 shows the results averaged over two rounds of active annotation and learning for five strategies with two types of model updates. While the performance for dissonance class across all strategies do not seem to boost much in a single round of active learning (since adding 300 new annotations adds only between 10-30 dissonance examples in each round), Figure 3 shows that the CM approach always performs better than IT. IT could help models generalize to new domains during transfer learning, but it may not add a lot of value when data is collected in the same domain in each iteration of the AL loop. This could be because IT biases the model towards the distribution of the latest sample set due to the effects of catastrophic forgetting (Yogatama et al., 2019) while CM implicitly balances all batches of data.

The performance of RANDOM-CM strategy lags behind the rest of the CM strategies. The other strategies perform better than RANDOM but one strategy does not offer significant advantages over another, further confirming the observation from Table 3 that the AL strategies have a significant overlap and could be choosing very similar samples.

5.3 Qualitative Evaluation of Annotation Costs

Table 5 displays the results of a study on the quality of annotation, measuring subjective difficulty and time taken. We sampled 300 examples from a data pool of 3,000 unannotated examples for each strategy so that the experiment is consistent with the unlabeled pool size used across other experiments for each of the strategies. Of these 300, we picked 125 (from each strategy) to get annotated for their difficulty on a scale of 0-5. This number was chosen based on balancing having enough examples per strategy for meaningful statistics while not taking too much of annotator's time and effort. The annotations were conducted on a simple annotation app that records the time taken to produce the first label an annotator decides on (i.e., any corrections to the label wouldn't count towards the time calculation). The Pearson correlation between the average time taken and the average difficulty value was 0.41.

	Rare %	Time (s)	Subj. diff.
RANDOM	3.20	11.96	-0.065
ENTROPY	6.80	12.78	0.035
CORESET	6.00	11.89	0.039
CAL	4.80	11.88	-0.045
PRC	**7.60**	**13.55**	**0.071**

Table 5: Evaluation of annotation difficulty by selection strategy. Rare % is how much the rare class (dissonance) was selected; Time is per instance and subj diff is z-scored subjective rating of difficulty. Our PRC approach selects the most rare class instances but also results in more costly annotations in terms of time and subjective ratings.

Annotation cost (in terms of time taken to annotate) is known to increase when employing active learning strategies compared to that of a random baseline (Settles et al., 2008). We find that PRC picks out the "most difficult" samples, and takes almost a second longer to annotate than average (average time taken: 12.59s), followed by ENTROPY and CORESET strategies – this complies with ENTROPY picking the most uncertain samples and CORESET executing diversity sampling and representing the data better, thus increasing the number of dissonance samples. The subjective difficulty reported is the average z-score of difficulty scores picked by the annotators. This is done to normalize the variability of subjective ratings. The inter-rater reliability for the entire exercise was measured using the Cohen's κ for two annotators, which was calculated to be 0.37 (fair agreement), with an overlap of 66%.

	F1-macro	F1-Dis	Prec-Dis	Rec-Dis	AUC
model_{iter0}	0.623	0.332	0.364	0.306	0.634
Deb→Sm	0.667	0.419	0.510	**0.355**	0.702
CE→Deb→Sm	0.658	0.389	0.483	0.327	0.707
Deb→Big	0.647	0.417	**0.695**	0.298	**0.753**
CE→Deb→Big	**0.669**	**0.425**	0.536	0.352	0.711

Table 6: The final dataset tested on the best transfer models from Tables 1 and 2 with CM approach. These models could subsequently be used to obtain newer samples more efficiently. model_{iter0} refers to the best model from continued training on Table 2, with scores reported on the final test and dev sets.

In general, we found that PRC addresses the rare-class challenge better than the other AL strategies. On transferring from CE/Debate corpora, the model is able to pick up on cues that indicate "Contrast" or "Disagreement" between two inputs, so PRC initially might pick samples with dissonant language (including cognitive and non-cognitive dissonance) with a high false positive rate, and improve over iterations. We also found that both the ENTROPY and CORESET strategies substantially increase the number of dissonant examples, thus partially addressing the needle-in-haystack problem.

5.4 A final dataset: Putting it all together.

We release two versions of train data: small and big; along with the development and test data (see §4.1) The *small* set comprises the 2,924 examples which were use for the active learning experiments discussed previously. Building on our learnings from the active learning experiments, we created a second (*big*) data set with 6,649 examples that includes the small plus an additional 3,725 examples derived over more rounds of active learning restricted to the PRC or ENTROPY strategies. It contains 692

dissonant samples, comprising 10.40% among all. Table 6 reports the performance improvement from using this final larger dataset, yielding the best performance so far with AUC > 0.75.

6 Conclusion

In this work, we have systematically studied approaches to key steps of active learning for tackling a rare-class modeling using a modern large language-modeling approach. While transformer-based systems have enabled greater accuracies with fewer training examples, data acquisition obstacles still persist for rare-class tasks – when the class label is very infrequent (e.g., < 5% of samples). We examined pool-based active annotation and learning in a real-world, rare class, natural language setting by exploring five common acquisition strategies with two different model update approaches. We found that a relatively simple acquisition using the probability of rare class for a model could lead to significant improvement in the rare class samples. We also qualitatively analyzed the data samples extracted from each data acquisition strategy by using subjective scoring and timing the annotators, finding PRC to be the most difficult to annotate, while also remaining the best method to improve rare class samples and model performance. Our final dataset of 9,589 examples (*Big* train + dev + test) is made available along with an implementation of the PRC method and our state-of-the-art model for cognitive dissonance detection.

7 Limitations

We use RoBERTa-base models trained on a single 12GB memory GPU (we used a NVIDIA Titan XP graphics card) for our experiments. Obtaining annotations for cognitive dissonance are limited by the availability of annotators and is not easily scalable in crowdsourcing platforms due to the required training and expertise in identifying dissonance. Due to this limitation, only two iterations of the AL loop for each setting were feasible for experiments. The transfer learning experiments in this paper were limited to two similar tasks, but there might be other tasks that could further improve or exceed the zero-shot performance of the models to cold start the active learning.

We focus on fine-tuning and active learning selection strategies to improve performance of rare-class classification for a specific task: dissonance detection across discourse units. Therefore, further work would be necessary to determine if the findings extend to other tasks. Additionally, the results may be different for other languages or time intervals of data collection. The performance of the neural parser on splitting tweets into discourse units can produce parses that are imperfect but the annotators and our systems worked off its output regardless to keep the process consistent. An improved discourse parser may also lead to improved annotator agreement and/or classifier accuracy. The dataset that we release from this paper, which contains labels of expressions of some cognitive states, was constructed using criteria that may not be fully objective.

8 Ethics Statement

The dataset for annotation was created from public social media posts with all usernames, phone numbers, addresses, and URLs removed. The research was approved by an academic institutional ethics review board. All of our work was restricted to document-level information; no user-level information was used. According to Twitter User Agreement, no further user content is required to use the publicly available data.

The detection of dissonance has many beneficial applications such as understanding belief trends and study of mental health from consenting individuals. However, it also could be used toward manipulative goals via targeted messaging to influence beliefs potential without users' awareness of such goals, a use-case that this work does not intend. Further, while we hope such models could be used to help better understand and assess mental health, clinical evaluations would need to be conducted before our models are integrated into any mental health practice.

Acknowledgements

We thank Lucie Flek (Data Science & Language Technologies, University of Bonn) and Ji-Ung Lee (UKP Lab, TU Darmstadt) for their insightful feedback about this work.

This work was supported by DARPA via Young Faculty Award grant #W911NF-20-1-0306 to H. Andrew Schwartz at Stony Brook University; the conclusions and opinions expressed are attributable only to the authors and should not be construed as those of DARPA or the U.S. Department of Defense.

References

Samir Al-Stouhi and Chandan K Reddy. 2016. Transfer learning for class imbalance problems with inadequate data. *Knowledge and information systems*, 48(1):201–228.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*.

Simon Chapman, Wai Leng Wong, and Wayne Smith. 1993. Self-exempting beliefs about smoking and health: differences between smokers and ex-smokers. *American journal of public health*, 83(2):215–219.

Jongwon Choi, Kwang Moo Yi, Jihoon Kim, Jinho Choo, Byoungjip Kim, Jinyeop Chang, Youngjune Gwon, and Hyung Jin Chang. 2021. Vab-al: Incorporating class imbalance and difficulty with variational bayes for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6749–6758.

Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. 2021. Batch active learning at scale. In *Advances in Neural Information Processing Systems*, volume 34, pages 11933–11944. Curran Associates, Inc.

Anja Dalgaard-Nielsen. 2013. Promoting exit from violent extremism: Themes and approaches. *Studies in Conflict & Terrorism*, 36(2):99–115.

Oscar Day and Taghi M Khoshgoftaar. 2017. A survey on heterogeneous transfer learning. *Journal of Big Data*, 4(1):1–42.

Melanie Ducoffe and Frederic Precioso. 2018. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Linton C Freeman. 1965. *Elementary applied statistics: for students in behavioral science*. New York: Wiley.

Daniel Gissin and Shai Shalev-Shwartz. 2019. Discriminative active learning. *arXiv preprint arXiv:1907.06347*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Tom SF Haines and Tao Xiang. 2014. Active rare class discovery and classification using dirichlet processes. *International Journal of Computer Vision*, 106(3):315–331.

Eddie Harmon-Jones and Cindy Harmon-Jones. 2007. Cognitive dissonance theory after 50 years of development. *Zeitschrift für Sozialpsychologie*, 38(1):7–16.

Eddie Harmon-Jones, Cindy Harmon-Jones, Meghan Fearn, Jonathan D Sigelman, and Peter Johnson. 2008. Left frontal cortical activation and spreading of alternatives: tests of the action-based model of dissonance. *Journal of personality and social psychology*, 94(1):1.

Eddie Harmon-Jones and Judson Mills. 2019. An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. *Cognitive dissonance: Reexamining a pivotal theory in psychology*.

Jason S Hartford, Kevin Leyton-Brown, Hadas Raviv, Dan Padnos, Shahar Lev, and Barak Lenz. 2020. Exemplar guided active learning. *Advances in Neural Information Processing Systems*, 33:13163–13173.

Timothy M. Hospedales, Shaogang Gong, and Tao Xiang. 2013. Finding rare classes: Active learning with generative and discriminative models. *IEEE Transactions on Knowledge and Data Engineering*, 25(2):374–386.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Swanie Juhng, Matthew Matero, Vasudha Varadarajan, Johannes Eichstaedt, Adithya V Ganesan, and H Andrew Schwartz. 2023. Discourse-level representations can improve prediction of degree of anxiety. In *Proceedings of The 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

David Kale and Yan Liu. 2013. Accelerating active learning with transfer learning. In *2013 IEEE 13th International Conference on Data Mining*, pages 1085–1090. IEEE.

Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher Manning. 2021. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. 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 7265–7281, Online. Association for Computational Linguistics.

Suraj Kothawade, Nathan Beck, Krishnateja Killamsetty, and Rishabh Iyer. 2021. Similar: Submodular information measures based active learning in realistic scenarios. In *Advances in Neural Information*

Processing Systems, volume 34, pages 18685–18697. Curran Associates, Inc.

Shoushan Li, Shengfeng Ju, Guodong Zhou, and Xiaojun Li. 2012. Active learning for imbalanced sentiment classification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 139–148, Jeju Island, Korea. Association for Computational Linguistics.

Yichan Liang, Jianheng Li, and Jian Yin. 2019. A new multi-choice reading comprehension dataset for curriculum learning. In *Asian Conference on Machine Learning*, pages 742–757. PMLR.

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 arXiv:1907.11692*.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

April McGrath. 2017. Dealing with dissonance: A review of cognitive dissonance reduction. *Social and Personality Psychology Compass*, 11(12):e12362.

Prateek Munjal, Nasir Hayat, Munawar Hayat, Jamshid Sourati, and Shadab Khan. 2022. Towards robust and reproducible active learning using neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 223–232.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*.

Andrew Ng. 2021. Mlops: from model-centric to data-centric ai. *Online unter https://www. deeplearning. ai/wp-content/uploads/2021/06/MLOps-From-Model-centric-to-Data-centricAI. pdf [Zugriffam09. 09.2021] Search in*.

Livia Polanyi. 1988. A formal model of the structure of discourse. *Journal of pragmatics*, 12(5-6):601–638.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*, volume 1. Vancouver, CA:.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*.

Youngseo Son, Vasudha Varadarajan, and H. Andrew Schwartz. 2022. Discourse relation embeddings: Representing the relations between discourse segments in social media. In *Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)*. Association for Computational Linguistics.

Maite Taboada and William C Mann. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse studies*, 8(3):423–459.

Akim Tsvigun, Artem Shelmanov, Gleb Kuzmin, Leonid Sanochkin, Daniil Larionov, Gleb Gusev, Manvel Avetisian, and Leonid Zhukov. 2022. Towards computationally feasible deep active learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1198–1218, Seattle, United States. Association for Computational Linguistics.

Vasudha Varadarajan, Nikita Soni, Weixi Wang, Christian Luhmann, H. Andrew Schwartz, and Naoya Inoue. 2022. Detecting dissonant stance in social media: The role of topic exposure. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*. Association for Computational Linguistics.

Dan Wang and Yi Shang. 2014. A new active labeling method for deep learning. *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 112–119.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967, Brussels, Belgium. Association for Computational Linguistics.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics.

Xueying Zhan, Qingzhong Wang, Kuan-hao Huang, Haoyi Xiong, Dejing Dou, and Antoni B Chan. 2022. A comparative survey of deep active learning. *arXiv preprint arXiv:2203.13450*.

Mike Zhang and Barbara Plank. 2021. Cartography active learning. *arXiv preprint arXiv:2109.04282*.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

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

B ☑ Did you use or create scientific artifacts?

4.1

☑ B1. Did you cite the creators of artifacts you used?
We are creating a data resource, and we cited all the datasets used in Section 4.1 and 4.2.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The dataset created is not yet available for distribution and we separately shared a anonymized dataset with the reviewers.

☑ 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)?
8

☑ 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?
8

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
8

☑ 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.
4

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

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?
4

☑ 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?
5

☒ 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.)?
Used RoBERTA-base from Huggingface, which is quite standard at this point in NLP and is self-explanatory in Section 4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
4

☑ 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.?
Figure 2. Since the task is subjective and the annotators label what they perceived as "dissonance" in language, there are no direct risks. General risks with the dataset in Section 8.

☒ 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)?
We recruited graduate students within the university with a background in psychology, paid a standard hourly rate for students at the university.

☐ 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?
Not applicable. Twitter's terms are mentioned in Section 8.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
8

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
8