Few-Shot (Dis)Agreement Identification in Online Discussions with Regularized and Augmented Meta-Learning

Yuanyuan Lei and Ruihong Huang
Department of Computer Science and Engineering
Texas A&M University, College Station, TX
{yuanyuan, huangrh}@tamu.edu

Abstract

Online discussions are abundant with opinions towards a common topic, and identifying (dis)agreement between a pair of comments enables many opinion mining applications. Realizing the increasing needs to analyze opinions for emergent new topics that however tend to lack annotations, we present the first meta-learning approach for *few-shot* (dis)agreement identification that can be quickly applied to analyze opinions for new topics with few labeled instances. Furthermore, we enhance the meta-learner's domain generalization ability from two perspectives. The first is domain-invariant regularization, where we design a lexicon-based regularization loss to enable the meta-learner to learn domain-invariant cues. The second is domain-aware augmentation, where we propose domain-aware task augmentation for meta-training to learn domain-specific expressions. In addition to using an existing dataset, we also evaluate our approach on two very recent new topics, mask mandate and COVID vaccine, using our newly annotated datasets containing 1.5k and 1.4k SubReddits comment pairs respectively[1]. Extensive experiments on three domains/topics demonstrate the effectiveness of our meta-learning approach[2].

1 Introduction

As seen in many online forums and SubReddits, people express different opinions and perspectives towards a common topic in online discussions, by posting their comments towards a given question or replying directly to another user's previous comments. Generally, we call the comments replying to another comments as *Response* and the comments being replied as *Quote* (Walker et al., 2012). Detecting agreement and disagreement relations between *(Quote, Response)* pairs addressing a shared topic will enable many opinion mining applications and inform policy making. However, realizing that new topics keep emerging, it is unrealistic to expect existing annotated datasets to cover every topics of interest. To avoid the time-consuming process to create a large annotated dataset for a new topic, we study *few-shot* agreement and disagreement identification that aims to quickly build a model on a new topic domain with few labeled instances.

To tackle the difficulty of *few-shot* agreement and disagreement identification under a new topic domain, we present a metric-based meta-learning approach that trains a meta-learner with two key abilities: deriving the class embedding from few provided support examples, and comparing the relation between new instance and each class embedding to make the prediction. Specifically, the meta-learner is trained on annotation-rich domains with the technique of episodic training (Vinyals et al., 2016): in each training episode, K examples per class are sampled as support set and N examples as query set, each query instance is compared with the class embeddings derived from the support set and the loss is minimized on the query set. Thus, when adapting to a new domain with very few labeled instances, the meta-learner has the ability of deriving the class embedding on the new domain from the provided few examples, and comparing a test instance with each class embedding via a learner relation network (Sung et al., 2018).

Inspired by prior research (Misra and Walker, 2013) that studied rich domain-independent indicators of agreement and denial in online discussions, we further encourage the meta-learning system to learn domain-invariant features and thus enhance its ability of quickly generalizing to a new test domain. Specifically, guided by (Misra and Walker, 2013), we complied a lexicon of domain-independent (dis)agreement indicators consisting of several hundreds of words or short phrases, e.g., indicators expressing agreement like "yes",

[1]The newly annotated SubReddit datasets: https://github.com/yuanyuanlei-nlp/SubReddit_agreement_dataset

[2]The code link: https://github.com/yuanyuanlei-nlp/fewshot_agreement_emnlp_2022

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5610 - 5622

"make sense", or conveying denial like "no", "but" etc, which can help our human beings distinguish (dis)agreement relation. Furthermore, we designed a *regularization loss* based on the lexicon and added it to the meta-learning system, so that the meta-learner pay more attention to the domain-independent cues. The designed lexicon-guided regularization loss can incorporate human knowledge into the meta-learner, encourage the meta-learner to focus more on domain-invariant features, so as to enhance its domain generalization ability.

Meanwhile, we design domain-aware task augmentation which decomposes the entire training dataset into clusters for episodic training, with each cluster corresponding to a topic domain, in order to better train the meta-learner to recognize domain-specific expressions of agreement and disagreements. Existing labeled datasets for agreement and disagreement identification usually contain data instances from multiple topic domains. If we randomly sample from an entire dataset for each meta-training episode, many episodes have sampled instances in the support set and the query set that do not match in domain and have divergent data distributions. The domain mismatch will lead to poor transfer between support and query sets (Murty et al., 2021), and thus hinders the meta-learner from learning to recognize domain-specific expressions of agreement and disagreements. Therefore, we perform domain-aware task augmentation for meta-training to strengthen the few-shot adaptation ability of the meta-learner, where we sample instances from the same domain to form support set and query set for each episode.

In addition to using an existing dataset for evaluation, we also evaluate our approach on very recent new topics. We collected and annotated 1.5k comment pairs from SubReddit on the mask mandate topic [3], as well as 1.4k comment pairs from SubReddit on the COVID vaccine topic[4] in the year of 2021. Experiments of both binary and three-class classification on the three domains/topics, shows that compared to the conventional fine tuning method and prompt-tuning method, the meta-learning approach achieves consistent and noticeable performance gains across the three domains under the challenging *few-shot* setting for (dis)agreement identification. Both of the two strategies for strengthening the adaptation ability of the meta-learner further improve the performance of the meta-learner, by enabling it to learn both domain-invariant cues and domain-specific expressions for (dis)agreement identification.

To summarize, our contributions are as follows:

- We present the first meta-learning approach for *few-shot* (dis)agreement identification.
- We design a lexicon based regularization loss to encourage the meta-learner to learn domain-invariant features.
- We perform domain-aware task augmentation to better train the meta-learner to recognize domain-specific expressions.

2 Related Work

(Dis)agreement identification research in online conversations or social media dialogues attracted increasing attentions. (Walker et al., 2012) provided the Internet Argument Corpus (IAC), annotating agreement/disagreement relation for Q-R (Quote-Response) post pairs in ten different domains. (Misra and Walker, 2013) conducted binary classification (*agreement vs. disagreement*) on the IAC corpus and studied rich domain independent cues for (dis)agreement identification. (Wang and Cardie, 2014) proposed to improve three-way classification (*agreement vs. neutral vs. disagreement*) with a socially-tuned sentiment lexicon. (Rosenthal and McKeown, 2015) introduced a larger dataset, the Agreement by Create Debaters (ABCD) corpus, and conducted three-way classification with transfer learning. However, none of the prior research has studied the (dis)agreement identification task under the cross-domain few-shot setting.

Meta-Learning has been studied for years for few-shot learning. Metric-based meta-learning learns an embedding function mapping individual instances into a representation space and learns a similarity function to calculate distance between two instances. Several metric-based meta-learners have been proposed, including Siamese Network (Koch et al., 2015), Matching Network (Vinyals et al., 2016), Prototype Network (Snell et al., 2017) and Relation Network (Sung et al., 2018). Another direction is optimization-based meta-learning (Finn et al., 2017) that aims to learn a good initialization to make a neural model reach the optimal for a new task quickly. We focus on developing a

[3] https://www.reddit.com/r/antimaskers/

[4] https://www.reddit.com/r/CovidVaccinated/

metric-based meta-learning model on the basis of Prototype and Relation Network models.

Meta-Learning for NLP has been studied for many NLP tasks under the few-shot setting, including topic classification (Jiang et al., 2018), entity relation classification (Sun et al., 2019; Geng et al., 2019), word sense disambiguation (Deng et al., 2020) and event detection (Deng et al., 2020; Lai et al., 2020). Mostly, prior works used meta-learning to identify unseen new classes and treat a class as a task, however, we aim to identify (dis)agreement in unseen new domains and treat a domain as a task (Yin, 2020).

Domain generalization has been studied long before the emergence of meta-learning, aiming to generalize from a set of seen domains to unseen domains without accessing any instance from the unseen domain during the training stage. As a strategy to achieve domain generalization, (Blanchard et al., 2011; Li et al., 2018; Muandet et al., 2013) proposed extracting domain-invariant features from various seen domains to enhance generalization ability. To the best of our knowledge, we lead on using domain-invariant features together with meta-learning to enhance few-shot generalization ability across domains.

Task augmentation for meta-learning was first studied in (Rajendran et al., 2020). Lack of well-defined data distribution is a recognized obstacle of meta-learning for solving NLP problems, generating attempts in augmenting meta-training tasks. (Bansal et al., 2020) proposed the SMLMT method to create new self-supervised tasks. Most closely related to our work is the strategy mentioned in (Murty et al., 2021) which clustered the entire dataset into several clusters by K-means and sampled support & query set from the same cluster to form training tasks. Our idea of task augmentation is different from theirs in that we relied on domain information to decompose the entire dataset into different training domains, creating clearer boundaries for different types of tasks.

3 The Meta-Learning Approach

In this section, we will elaborate our meta-learning approach in details. Firstly, we introduced the structure of the basic meta-learning model. Then we enhanced the model's domain generalization ability from two perspectives: (1) Manually created a lexicon for *domain-independent* (dis)agreement indicators, and designed a regularization loss to make the meta-learner focus more on domain-invariant features, (2) Decomposed the entire training dataset into several sub-datasets based on *domain-specific* info to augment the task distribution. Fig. 1 illustrated the pipeline of our meta-learning approach.

3.1 The Basic Meta-Learning Model

In the cross-domain few-shot (dis)agreement identification problem, we are given a training dataset $\mathcal{D}_{meta-train}$ consisting of rich labeled Q-R pairs from various domains, and a testing dataset $\mathcal{D}_{meta-test}$ in an unseen new domain. $\mathcal{D}_{meta-test}$ is splitted into two parts: a support set $\mathcal{D}_{test-support}$ with only a small number of K labeled Q-R pairs per class, and a test set $\mathcal{D}_{test-query}$ used to evaluate the model performance on. Our goal is to train a meta-learner $f : (S, x) \rightarrow \hat{y}$ that takes a support set $S = \{s_k^i, i \in 1 \dots C, k \in 1 \dots K\}$ and a test instance x from $\mathcal{D}_{test-query}$ as input, and returns a prediction $\hat{y}$ for the instance (x, y), where $y \in \{1, \dots, C\}$ is the true label and C is the number of classes. Specifically, $C = 2$ in the two-way classification setting, where we classify the agreement or disagreement relation between a *Quote* comment and a *Response* comment; $C = 3$ in the three-way classification setting, where we classify the agreement, neutral, or disagreement relation between a *Quote* and a *Response*. K is the number of labeled support examples for each class, and the support set S has $C * K$ examples in total. The few-shot problem is often named a C-way K-shot learning problem.

3.1.1 Episodic Training

To mimic the meta-testing task that takes a support set $\mathcal{D}_{test-support}$ & test instances $\mathcal{D}_{test-query}$ as input, and make the model accustomed to the few-shot environment, we followed the episodic training idea in (Vinyals et al., 2016) to create training tasks: randomly sampled K labeled examples per class from the training dataset as the support set $\mathcal{D}_{train-support}$ ($K * C$ support examples in total), and N query examples from the rest of training data as query set $\mathcal{D}_{train-query}$, output prediction values for query examples and minimized the loss on the query set to update the meta-learning model. Note that the K labeled support examples in the testing dataset $\mathcal{D}_{test-support}$ did not participate in the training stage, but just served as model input in testing tasks.

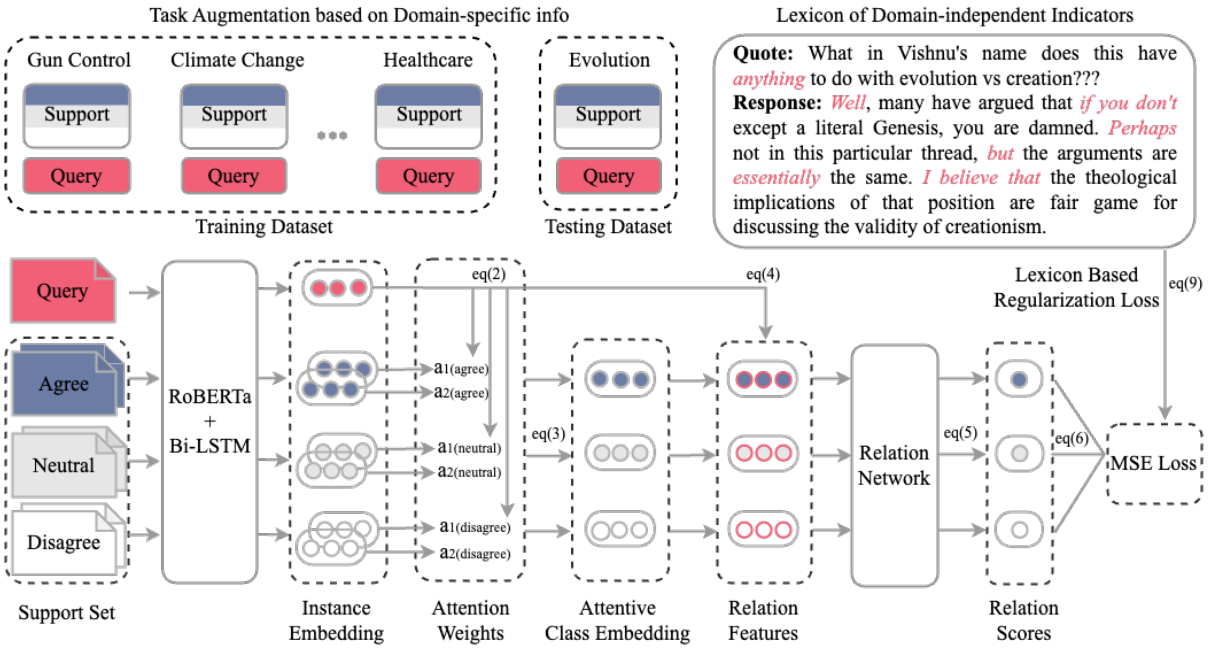

Figure 1: Illustration of the meta-learning approach for 3-way 2-shot problem with one query instance

3.1.2 Attentive Class Embedding Building

Within a training/testing task, each class embedding is derived attentively from the given support examples via learned attention weights on them: first obtained the Q-R pair embedding for each support & query sample, then mapped support examples through two-layer neural networks learned separately for each class, lastly calculated the attention weights to derive attentive class embedding.

The initial embedding for the support $s_k^i, i \in \{1, \ldots, C\}, k \in \{1, \ldots, K\}$ and query examples e_q are obtained by using pre-trained RoBERTa model (Liu et al., 2019) with an additional Bi-LSTM layer added (Hochreiter and Schmidhuber, 1997) on top. We concatenate the hidden state vectors of quote and response sentences at their sentence start token <s> and use the concatenation as the pair embedding.

Then, we mapped support examples through a two-layer neural network learned separately for each class:

$$\hat{s}_k^i = W_2^i(W_1^i s_k^i + b_1^i) + b_2^i \quad (1)$$

where $i \in \{1 \ldots C\}, k \in \{1 \ldots K\}$, W_1^i, W_2^i, b_1^i, b_2^i are the matrices parameters in the two-layer neural networks that map support examples for each class.

At last, support examples were aggregated into class embedding via learned attentions $\{a_k^i\}_{k=1}^K$ over $\{\hat{s}_k^i\}_{k=1}^K$, in which attention weights are calculated wrt both support $\{\hat{s}_k^i\}$ and query e_q:

$$a_k^i = softmax(m^T ReLU(W_3 \hat{s}_k^i + W_4 e_q)) \quad (2)$$

$$c_i = \sum_{k=1}^{K} a_k^i * \hat{s}_k^i \quad (3)$$

where $c_i, i \in \{1, \ldots, C\}$ is i-th class embedding, W_3, W_4, m are the matrices parameters in the neural network that learns attention weights.

Different from the naive mean average in the original Prototype Network (Snell et al., 2017), our method derived class embedding from support set in an attentive way, and also took the query instance into consideration when calculating the attention weight over support examples. Intuitively, a support instance has higher attention weight in deriving the class embedding if it is closer to the query in the representation space.

3.1.3 Relation Network

With the classes embedding and query embedding at hand, the final step is to compare the query instance with each class embedding via a learned two-layer relation network, output the relation scores for each class, and choose the class with maximum relation score as the prediction result.

For each class, relation feature is designed as the concatenation of class embedding c_i, query embed-

ding e_q, and the element-wise subtraction, element-wise multiplication, L2 norm, dot product of them:

$$f_{iq} = [c_i; e_q; c_i - e_q; c_i \odot e_q; ||c_i - e_q||; c_i \cdot e_q] \quad (4)$$

Then relation features were fed into a two-layer relation network to learn the relation scores between the i-th class and query e_q as output:

$$r_{iq} = sigmoid(W_6(W_5 f_{iq} + b_5) + b_6) \quad (5)$$

where W_5, W_6, b_5, b_6 are the matrices parameters in the two-layer relation network.

Output relation score r_{iq} is a scalar between 0 and 1 to measure the similarity between query instance and each class, and the ground truth $y_q \in \{0, 1\}$ meaning matched class has similarity 1 & mismatched class has similarity 0. The objective function we used is the mean square error (MSE) loss on the query set:

$$L_{MSE} = \sum_{i=1}^{C} \sum_{q=1}^{N} (r_{iq} - I(y_q == i))^2 \quad (6)$$

3.2 Lexicon Based Regularization Loss

To further strengthen model's ability of quickly generalizing to a new domain, we manually created a lexicon of domain-independent (dis)agreement indicators to incorporate domain-invariant features from various seen domain. Moreover, we designed a *lexicon based regularization loss* to make the meta-learner focus more on selected domain-independent indicators.

When creating the domain-independent lexicon, we followed the similar scenario in prior work (Misra and Walker, 2013), which proved domain-independent words/phrases in cue words, agreement words, denial words, and hedge words categories are all crucial to cross-domain (dis)agreement identification. We manually inspected the development set in our experimental datasets, and selected the words/phrases belonging to discourse markers associated with stating a personal opinion (cue words), agreement markers expressing support (agreement words), denial markers showing rejection/negation (denial words), and hedges that deliberately vague/soften a claim (hedge words), which are important for human to identify agree/disagree. Besides, in order to provide better generalization, we generalized the selected phrases, e.g., *I don't think* would also result in *I don't see* being added into the lexicon (Misra and Walker, 2013). Table 1 listed examples of our selected domain-independent words/phrases (Please refer to Appendix A.3 for the full lexicon).

Category (number)	Examples
Cue words (48)	so, oh, well, just, and, because, though, as well, if, then, thus, unless, seems, also, you, uh
Agreement (145)	yes, correct, agree, accept, support, true, like, good, exactly, ok, right, clear, sure, thanks, believe, of course, make sense
Denial (278)	no, not, never, nothing, however, but, doesn't, don't, isn't, yet, none, hate, false, wrong, doubt, disagree, how can, I don't think
Hedge (25)	maybe, probably, would, could, rather, although, really, actually, wondering, possibly, essentially, anyway, somewhat, I suppose

Table 1: Examples of selected domain-invariant features

To make the meta-learner focus more on the domain-independent features, we designed a *regularization loss* to maximize the model's attention on selected domain-invariant words. For an instance consisting of n words, the model's attention on the l^{th} word is designed as L2 norm of the gradient of model output (relation scores) wrt l^{th} word's embedding. Thus, model's attention on the words in a query instance e_q is:

$$\overrightarrow{g_q} = (||\frac{\partial r_{tq}}{\partial w_{1q}}||, ||\frac{\partial r_{tq}}{\partial w_{2q}}||, \ldots, ||\frac{\partial r_{tq}}{\partial w_{nq}}||) \quad (7)$$

where $(w_{1q}, w_{2q}, \ldots, w_{nq}) \in e_q$ and $y_q = t$. Similarly, the attention on the words in a support example s_k^t is:

$$\overrightarrow{g_{s_k^t}} = (||\frac{\partial r_{tq}}{\partial w_{1s_k^t}}||, ||\frac{\partial r_{tq}}{\partial w_{2s_k^t}}||, \ldots, ||\frac{\partial r_{tq}}{\partial w_{ns_k^t}}||) \quad (8)$$

where $(w_{1s_k^t}, w_{2s_k^t}, \ldots, w_{ns_k^t}) \in s_k^t$. Bigger gradient value means more influence on the model output, and thus means more model attention. Then, we used an indicator $I(w_1, w_2, \ldots, w_n)$ to show whether the word belongs to our selected domain-independent words set, which is a vector consisting of value 0 or 1. Finally, our *regularization loss* is designed as the dot product of gradient vector

(model attention) and indicator vector:

$$Lreg = -\sum_{q=1}^{N} \Big\{ \overrightarrow{g_q} \cdot I(w_{1q}, w_{2q}, \ldots, w_{nq}) + \sum_{k=1}^{K} \overrightarrow{g_{s_k^t}} \cdot I(w_{1s_k^t}, w_{2s_k^t}, \ldots, w_{ns_k^t}) \Big\} \quad (9)$$

where $y_q = t$. The total objective loss will be:

$$L_{total} = L_{MSE} + \lambda * Lreg \quad (10)$$

where λ is a hyper-parameter.

3.3 Meta-training Task Augmentation

In our previous episodic training process, we treat the entire training dataset as tasks, meaning support set and query set are sampled from the entire training dataset for each single training task, which is also the common approach used by previous papers (Murty et al., 2021). This brings us two major problems: one is meta-learning actually needs a well-defined task distribution from which a large number of diverse tasks can be sampled to train the meta-learner, another one is the entire training dataset consisting of various domains data is also heterogeneous. Thus, sampling support & query from the entire training dataset not only limited the diversity of meta-training tasks, but also resulted in support & query examples are heterogeneous with each other, making the meta-learner harder to foster the ability of quickly adapting to new domains.

For these reasons, we proposed to augment the task distribution by decomposing the entire training dataset into several sub-datasets based on domain-specific information and sampling support & query from the same sub-dataset to form training tasks. To be detailed, for the dataset having ground-truth domain labels, we divided the dataset by golden domain labels. But for the dataset with no domain labels, we decomposed the dataset by clustering discussion thread titles which usually indicate domain or the topic under discussion. The K-means algorithm (MacQueen, 1967) is applied on the embedding of each discussion title's start token <s> obtained from the pre-trained RoBERTa (Liu et al., 2019), and the number of clusters is selected by elbow method (Joshi and Nalwade, 2013). In this way, training tasks are created by sampling support & query from the same sub-datasets, and ideally the same training domain.

Dataset	Thread	Pairs	Agree	Neutral	Disagree
ABCD	10468	128343	28111	60128	40104
IAC	1806	9980	1113	2712	6155

Table 2: Statistics of training datasets

Testing dataset	Agree	Neutral	Disagree	Total
AWTP	740	985	757	2482
MaskMandate	645	343	546	1534
CovidVaccine	272	703	425	1400

Table 3: Statistics of testing datasets

4 Experiments

4.1 Datasets

Our experiments use five datasets, IAC and ABCD are used as training datasets, while AWTP, MaskMandate, and CovidVaccine are used as testing datasets, among which the latter two are new datasets created by ourselves (the annotation guidelines are in Appendix A.1.). Table 2-3 show the statistics for the training and testing datasets.

Internet Argument Corpus (IAC) (Walker et al., 2012) annotated post pairs from the website 4forums.com with scores from -5 to 5. We converted the average score into the label as previous paper did (Wang and Cardie, 2014)(Misra and Walker, 2013): [-5,-1] as *Disagreement*, (-1,1) as *Neutral*, [1,5] as *Agreement*. Also, pairs in IAC have human-annotated domain labels in a total of ten domains.

Agreement by Create Debaters (ABCD) dataset (Rosenthal and McKeown, 2015) collected post pairs from another website createdebate.com. Note the relation between two different users' posts can only be *Agreement* or *Disagreement* in ABCD.

Agreement in Wikipedia Talk Pages (AWTP) dataset (Andreas et al., 2012) collected Q-R pairs from LiveJournal Blogs and Wikipedia Edit Discussions, and annotated them directly with *Agreement*, *Neutral*, and *Disagreement* labels.

SubReddit-MaskMandate is a dataset we collected from a sub forum on reddit.com[5] with the topic of make mandate, discussing whether people should wear masks during the COVID pandemic. The Cohen's kappa between annotators is 0.8012.

SubReddit-CovidVaccine is a dataset we collected from a sub forum on reddit.com[6] with the topic of

[5]https://www.reddit.com/r/antimaskers/
[6]https://www.reddit.com/r/CovidVaccinated/

COVID vaccine, discussing whether people should take the COVID vaccination. The Cohen's kappa between annotators is 0.8215.

4.2 Experimental Settings

Experiments Design: In real life, there are discussion forums with sub-topics explicitly known, and also discussion forums without clear topic/domain labels. So we designed the experiments for both types of the discussions separately:

- **Training dataset with golden domain label**: We did experiments of both 2-way and 3-way classification using IAC as the training dataset $\mathcal{D}_{meta-train}$, and test on three domains/topics: AWTP, MaskMandate, CovidVaccine. When augmenting meta-training tasks, we divided IAC by its golden domain label (Table 4 5).
- **Training dataset with no domain label**: We did experiment using ABCD as training dataset $\mathcal{D}_{meta-train}$, and used the same three datasets as testing domains $\mathcal{D}_{meta-test}$. Since the relation between two different users' posts can only be *Agreement* or *Disagreement* in ABCD, we can only conduct 2-way classification generalizing from ABCD to testing domains. When augmenting meta-training tasks, we clustered the discussion titles in ABCD with K-means and selected five as the number of clusters using the elbow method (Table 6).

Implementation Details: The number of support examples per class K is set to 5, and the query set size N is set to 15 in each meta-training task. The λ in equation (10) is set to 1. We used AdamW (Loshchilov and Hutter, 2019) as the optimizer. The weight decay is set to be 1e-2. For 2-way and 3-way classification training on IAC, the numer of epochs is 2. For 2-way classification training on ABCD, the number of epoch is 1.

Evaluation Settings The testing task in a new domain consists of a support set $\mathcal{D}_{test-support}$ and a real test set $\mathcal{D}_{test-query}$ to evaluate the predictions on. We obtain $\mathcal{D}_{test-support}$ by randomly sampling $K = 5$ instances per class from a testing dataset, and then use the remaining data as $\mathcal{D}_{test-query}$ for evaluation. To control for variations due to $\mathcal{D}_{test-support}$, we repeat the sampling and evaluation procedure for 50 times, and report the average results. Here, we used F1 score for each class and macro Precision/Recall/F1 scores as evaluation metrics.

4.3 Baselines

Fine tuning and prompt-based tuning are the current common methods to solve few-shot problems, and we experimented with both methods as our baselines. The same as in our meta-learning model, RoBERTa (Liu et al., 2019) plus a Bi-LSTM layer on top is used to derive the embedding for support and query examples in the baselines.

- **RoBERTa training**: A classification layer is built on the concatenation of hidden state at sentence start token <s> of *Quote* and *Response* sentence. Then we trained this model on $\mathcal{D}_{meta-train}$ with the cross entropy loss.
- **Fine tuning**: We trained the above classification model on $\mathcal{D}_{meta-train}$ and then fine tuned on the few labeled examples (support set) of the testing domain $\mathcal{D}_{test-support}$.
- **Prompt training**: We designed eight prompt templates after referring to the commonly used templates mentioned in (Liu et al., 2021), and selected the best template based on the experiments of 2-way and 3-way classification generalizing from IAC to AWTP. The template "Q ?<mask>, R" with the label words "Yes, No, Maybe" is selected, where Q represents *Quote* sentences and R represents *Response* sentences, because it has the best performance compared to other templates (Appendix A.2 shows results for all the eight prompt templates). We inserted the selected template into the input text and predicted the word in the <mask> token, and trained the model on $\mathcal{D}_{meta-train}$.
- **Prompt tuning**: After training the above prompt-based model on $\mathcal{D}_{meta-train}$, we further tuned it on the few labeled examples (support set) of the testing domain $\mathcal{D}_{test-support}$, to learn the features of the new domain.

4.4 Results

Experimental results are summarized in Table 4-6. Results of the four baseline models are reported in the first four rows. The performance of basic meta-learning model (Meta), only adding the lexicon based regularization loss (Meta + reg), only with task augmentation (Meta + aug), and task augmentation together with regularization loss (Meta + aug + reg) are reported in the latter four rows.

Test Domain	AWTP					MaskMandate					CovidVaccine				
Model	A	D	Macro			A	D	Macro			A	D	Macro		
	F1		P	R	F1	F1		P	R	F1	F1		P	R	F1
RoBERTa training	76.4	70.5	74.9	73.7	73.4	68.8	73.5	72.0	71.4	71.1	76.3	71.6	75.1	74.2	73.9
Fine tuning	74.1	75.0	76.4	75.0	74.6	71.8	70.3	73.6	71.7	71.0	74.8	71.7	76.0	73.9	73.3
Prompt training	76.2	77.6	77.0	76.9	76.9	74.5	71.6	73.4	73.1	73.0	77.0	72.6	75.8	75.0	74.8
Prompt tuning	76.5	75.7	77.6	76.4	76.1	72.8	71.8	74.0	72.7	72.3	75.4	73.0	75.9	74.5	74.2
Meta	76.8	76.5	76.6	76.6	76.6	73.4	73.5	73.5	73.5	73.5	76.8	72.7	75.7	74.9	74.7
Meta + reg	79.1	76.6	78.3	77.9	77.9	74.8	73.8	74.3	74.3	74.3	77.0	74.4	76.1	75.8	75.7
Meta + aug	78.3	76.7	77.7	77.5	77.5	74.0	**74.4**	74.2	74.2	74.2	77.8	74.7	76.9	76.4	76.2
Meta + aug + reg	**79.1**	**79.8**	**79.5**	**79.4**	**79.4**	**77.8**	74.0	**76.7**	**76.0**	**75.9**	**79.3**	**75.5**	**78.6**	**77.6**	**77.4**

Table 4: Results of 2-way classification training on IAC (A: Agreement, D: Disagreement, P: Precision, R: Recall, F1: F1 score)

Test Domain	AWTP						MaskMandate						CovidVaccine					
Model	A	N	D	Macro			A	N	D	Macro			A	N	D	Macro		
	F1			P	R	F1	F1			P	R	F1	F1			P	R	F1
RoBERTa training	67.5	49.0	41.5	55.3	53.1	52.7	51.4	41.8	58.8	51.0	50.5	50.7	53.2	37.2	64.4	55.1	51.6	51.6
Fine tuning	62.3	41.9	51.4	55.7	53.0	51.9	51.3	37.1	58.6	52.0	49.6	49.0	54.2	39.7	64.7	54.2	53.4	52.9
Prompt training	65.2	41.1	50.8	53.1	53.5	52.4	51.7	39.2	59.1	53.0	51.2	50.0	54.1	37.9	59.2	52.7	50.6	50.4
Prompt tuning	64.4	45.3	49.5	57.0	54.3	53.1	52.1	40.5	59.7	50.8	50.9	50.8	53.8	44.5	60.5	55.8	53.5	52.9
Meta	67.5	43.4	49.0	53.9	54.5	53.3	**52.8**	41.5	59.9	52.4	51.2	51.4	59.7	38.4	64.9	55.3	55.2	54.3
Meta + reg	**68.8**	47.0	49.2	55.6	56.0	55.0	48.0	44.8	64.0	52.3	52.6	52.3	**60.0**	40.9	67.2	56.4	57.0	56.0
Meta + aug	67.4	47.5	48.6	55.5	55.0	54.5	50.2	43.4	64.0	52.5	52.8	52.5	56.1	46.8	66.8	**60.0**	55.8	56.6
Meta + aug + reg	68.0	**48.9**	**54.0**	**57.6**	**56.6**	**56.9**	46.0	**48.9**	**64.4**	**54.4**	**53.2**	**53.1**	56.2	**47.4**	**68.2**	57.8	**57.1**	**57.3**

Table 5: Results of 3-way classification training on IAC (A: Agreement, N: Neutral, D: Disagreement, P: Precision, R: Recall, F1: F1 score)

Comparing the row "Fine tuning" with the row "RoBERTA training" in Table 4-6, we can observe that fine tuning on the few labeled examples sometimes will lower the performance. As revealed in (Zhang et al., 2021), it is due to fine tuning's drawback of overfitting, making the model simply memorize the labeled samples and fail to learn generalizable features for new domains.

We can also see that prompt-based tuning method also faces several drawbacks. The performance of prompt tuning is sensitive to the template design (as shown in Appendix A.2), and slightly different templates can lead to noticeable performance drop. As analyzed in (Liu et al., 2021), human-involved heavy engineering for template design is required. Moreover, comparing "Prompt tuning" with "Prompt training" rows, we can observe that prompt-based tuning cannot make good use of the few labeled examples in new domain, leading to performance decrease sometimes.

Comparing Meta + aug + reg with baselines, our regularized and augmented meta-learning model can achieve noticeable improvement on both Precision and Recall on all three testing domains under both 2-way and 3-way classification. Precision can be increased by up to 2.8% compared to the best baseline performance, and Recall can be increased by up to 3.6%. The macro F1-score can be improved by 2.4% - 4.4%. Our meta-learning approach generalizes across domains well and do not require heavy engineering for template design.

4.5 Ablation Study

Effectiveness of lexicon based regularization loss Comparing Meta + reg with Meta, the results show making meta-learner focus more on domain-independent features through lexicon based regularization can generate better performance across all the testing domains, in both Precision and Recall.

Effectiveness of task augmentation Comparing Meta + aug with Meta, we can see that task augmentation will also yield improvements in both Precision and Recall. Not only dividing the training dataset by golden domain label (Table 4-5) can boost performance, but clustering can also improve the model (Table 6).

4.6 Effect of domain-invariant regularization

Training with regularization loss can indeed make the meta-learner pay more attention to the selected domain-independent words. The average of model attentions on our domain-independent words was increased by 0.4% to 1.2% across our experiments.

Moreover, the meta-learner will not degenerate

Test Domain	AWTP					MaskMandate					CovidVaccine				
Model	A	D	Macro			A	D	Macro			A	D	Macro		
	F1		P	R	F1	F1		P	R	F1	F1		P	R	F1
RoBERTa training	76.6	70.3	75.2	73.8	73.5	71.7	71.4	71.5	71.5	71.5	74.7	72.6	73.9	73.7	73.6
Fine tuning	77.3	74.0	76.7	75.9	75.7	72.3	67.7	72.6	70.6	70.0	74.7	73.2	74.8	74.1	73.9
Prompt training	76.5	71.0	75.1	74.0	73.7	73.1	70.1	71.9	71.6	71.6	76.1	73.7	75.2	75.0	74.9
Prompt tuning	75.8	75.6	76.5	75.9	75.7	70.8	69.2	71.4	70.4	70.0	76.4	74.9	76.3	75.8	75.6
Meta	77.6	75.2	76.7	76.4	76.4	73.4	70.1	72.1	71.8	71.8	76.6	74.3	75.8	75.6	75.5
Meta + reg	78.6	76.7	77.9	77.7	77.7	**76.2**	71.5	**74.8**	74.0	73.8	76.5	77.1	76.8	76.8	76.8
Meta + aug	77.1	77.0	77.1	77.1	77.1	76.0	71.7	74.6	74.0	73.8	**79.1**	74.7	**78.3**	77.2	76.9
Meta + aug + reg	**80.3**	**77.6**	**79.5**	**79.1**	**79.0**	74.1	**74.7**	74.5	**74.4**	**74.4**	78.4	**77.6**	78.1	**78.0**	**78.0**

Table 6: Results of 2-way classification training on ABCD (A: Agreement, D: Disagreement, P: Precision, R: Recall, F1: F1 score)

λ	0.0	0.25	0.50	0.75	1.0	1.5	2.0	3.0	5.0
Agree F1	78.3	78.2	79.1	77.8	79.1	79.2	79.3	78.7	79.5
Disagree F1	76.7	77.3	76.6	80.0	79.8	77.6	77.2	77.8	76.4
macro Precision	77.7	77.8	78.3	79.4	79.5	78.6	78.5	78.3	78.6
macro Recall	77.5	77.7	77.9	79.0	79.4	78.4	78.3	78.2	78.1
macro F1	77.5	77.7	77.9	78.9	79.4	78.4	78.2	78.2	78.0

Table 7: Performance change with different regularization weights

to perform classification only based on lexicon words after regularization. The model attention on non-lexicon words after adding the regularization loss is decreased only by 6% to 18% across different experiments. Majority attention has been retained for non-lexicon words, which meanwhile shows that domain-specific information is important for (dis)agreement identification.

We present the performance change across the different value of regularization weight λ in Table 7 for the experiment of 2-way classification that was trained on IAC and tested on AWTP domain. When λ is less than 1.0, the macro F1 increases as λ becomes bigger, and after λ exceeds 1.0, the macro F1 will decrease, but still outperforms the model without regularization ($\lambda = 0$).

4.7 Effect of support set size K

We study the effect of the number of support examples provided in the support set. K is the number of support examples for each class. The total size of the support set is $C * K$ where C is the number of classes. Table 8 presents the performance change across different value of K, for the experiment of 2-way ($C = 2$) classification that was trained on IAC and tested on AWTP.

We can see that when K is less than 5, the macro F1 keeps increasing as more support examples are provided. When K exceeds 5, the performance tends to level up and fluctuate a little.

K	1	2	3	4	5	10	15
macro F1	75.5	76.6	77.3	78.3	79.4	78.1	77.6
K	20	25	30	35	40	45	50
macro F1	78.9	77.9	77.1	79.4	78.7	78.7	78.4

Table 8: Performance change with different number of support examples per class.

5 Conclusion

In this paper, we developed a metric-based meta-learning model for few-shot (dis)agreement identification problem. Furthermore, we enhance the meta-learner's domain generalization ability from two perspectives: domain-invariant regularization through a lexicon-based regularization loss to learn domain-invariant features, and domain-aware task augmentation to learn domain-specific expressions.

6 Limitations

We are aware of two limitations of our meta-learning approach. First, while yielding better results, the model trained with the lexicon-based regularization loss is 2X slower in terms of training time. Second, since the regularization loss is a second-order gradient matrix, the model will require more memory in GPU during calculation.

Acknowledgements

We gratefully acknowledge support from National Science Foundation via the award IIS-1909252.

References

Jacob Andreas, Sara Rosenthal, and Kathleen McKeown. 2012. Annotating agreement and disagreement in threaded discussion. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 818–822, Istanbul, Turkey. European Language Resources Association (ELRA).

Trapit Bansal, Rishikesh Jha, Tsendsuren Munkhdalai, and Andrew McCallum. 2020. Self-supervised meta-learning for few-shot natural language classification tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 522–534.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. 2011. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Shumin Deng, Ningyu Zhang, Jiaojian Kang, Yichi Zhang, Wei Zhang, and Huajun Chen. 2020. Meta-learning with dynamic-memory-based prototypical network for few-shot event detection. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 151–159, New York, NY, USA. Association for Computing Machinery.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Ruiying Geng, Binhua Li, Yongbin Li, Yuxiao Ye, Ping Jian, and Jian Sun. 2019. Few-shot text classification with induction network. *CoRR*, abs/1902.10482.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Xiang Jiang, Mohammad Havaei, Gabriel Chartrand, Hassan Chouaib, Thomas Vincent, Andrew Jesson, Nicolas Chapados, and Stan Matwin. 2018. Attentive task-agnostic meta-learning for few-shot text classification.

Kalpana D. Joshi and Prakash S. Nalwade. 2013. Modified k-means for better initial cluster centres.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition.

Viet Dac Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2020. Exploiting the matching information in the support set for few shot event classification. *CoRR*, abs/2002.05295.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. 2018. Domain generalization with adversarial feature learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ArXiv*, abs/2107.13586.

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*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

J. B. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

Amita Misra and Marilyn Walker. 2013. Topic independent identification of agreement and disagreement in social media dialogue. In *Proceedings of the SIGDIAL 2013 Conference*, pages 41–50, Metz, France. Association for Computational Linguistics.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 10–18, Atlanta, Georgia, USA. PMLR.

Shikhar Murty, Tatsunori Hashimoto, and Christopher D Manning. 2021. Dreca: A general task augmentation strategy for few-shot natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1113–1125.

Janarthanan Rajendran, Alexander Irpan, and Eric Jang. 2020. Meta-learning requires meta-augmentation. In *Advances in Neural Information Processing Systems*, volume 33, pages 5705–5715. Curran Associates, Inc.

Sara Rosenthal and Kathleen McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 168–177.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4080–4090, Red Hook, NY, USA. Curran Associates Inc.

Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. 2019. Hierarchical attention prototypical networks for few-shot text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 476–485, Hong Kong, China. Association for Computational Linguistics.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29:3630–3638.

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).

Lu Wang and Claire Cardie. 2014. Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 97–106, Baltimore, Maryland. Association for Computational Linguistics.

Wenpeng Yin. 2020. Meta-learning for few-shot natural language processing: A survey. *CoRR*, abs/2007.09604.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115.

A Example Appendix

A.1 Annotation guideline for the MaskMandate and CovidVaccine datasets

Here, we describe our annotation guideline for the two datasets collect from the SubReddit. Every *(Quote, Response)* pairs are annotated with *Agreement*, *Disagreement*, or *Neutral* label:

- *Agreement*: both the *Quote* and *Response* express subjective opinions towards some topics instead of sharing objective experience or giving objective advice, and there is explicit or implicit evidence for strong agreement / support / semantically similarity.
- *Disagreement*: both *Quote* and *Response* are expressing subjective opinions, and there is explicit or implicit evidence for disagreement / attack / denial / reject / challenge / semantically opposition.
- *Neutral*: *Quote* or *Response* is not expressing subjective opinions, or *Quote* and *Response* are stating quite different / unrelated things, or both *Quote* and *Response* express opinions but *Response* is expressing vague / not sure / weak agreement or disagreement, or partially agree partially disagree.

A.2 Performance of eight prompt templates

We present the results for the eight prompt templates in the experiment of 2-way and 3-way classification training on IAC and testing on AWTP in Table 9 and Table 10. Comparing the eight prompt templates, the template "Q ?<mask>, R" with the label words "Yes, No, Maybe" corresponding to the *Agreement, Disagreement, Neutral* class respectively has the best performance, thus is selected as our template used for prompt-based baselines. Q represents *Quote* sentences and R represents *Response* sentences.

A.3 Domain-independent Lexicon

Here we released the full lexicon of domain-independent (dis)agreement indicators in Table 11.

Template	label words	Agree F1	Disagree F1	macro Precision	macro Recall	macro F1
Q ?<mask>, R	Agree, Disagree, Neutral	78.3	75.1	77.3	76.8	76.7
Q ?<mask>, R	Yes, No, Maybe	76.2	77.6	77.4	77.0	76.9
Q <mask>, R	Agree, Disagree, Neutral	72.8	72.8	76.5	73.7	72.8
Q <mask>, R	Yes, No, Maybe	77.9	71.7	76.8	75.2	74.8
Q ?<mask> R	Agree, Disagree, Neutral	76.8	68.4	75.7	73.3	72.6
Q ?<mask> R	Yes, No, Maybe	68.6	74.5	75.7	72.5	71.6
Q .<mask>, R	Agree, Disagree, Neutral	78.7	73.5	77.7	76.4	76.1
Q .<mask>, R	Yes, No, Maybe	78.3	75.6	77.0	76.9	76.9

Table 9: Performance of eight prompt templates for the experiment of 2-way classification training on IAC and testing on AWTP. Q represents *Quote* sentences and R represents *Response* sentences in templates.

Template	label words	Agree F1	Neutral F1	Disagree F1	macro Precision	macro Recall	macro F1
Q ?<mask>, R	Agree, Disagree, Neutral	67.4	46.3	45.0	55.1	54.1	52.9
Q ?<mask>, R	Yes, No, Maybe	64.4	45.3	49.5	57.0	54.3	53.1
Q <mask>, R	Agree, Disagree, Neutral	52.7	48.7	18.6	48.9	44.4	40.0
Q <mask>, R	Yes, No, Maybe	61.4	37.2	55.2	54.5	52.4	51.3
Q ?<mask> R	Agree, Disagree, Neutral	52.0	38.8	48.7	50.5	47.8	46.5
Q ?<mask> R	Yes, No, Maybe	62.9	40.3	50.5	52.1	52.3	51.2
Q .<mask>, R	Agree, Disagree, Neutral	65.6	38.9	53.5	55.3	54.0	52.7
Q .<mask>, R	Yes, No, Maybe	62.9	50.0	45.5	55.9	53.1	52.8

Table 10: Performance of eight prompt templates for the experiment of 3-way classification training on IAC and testing on AWTP. Q represents *Quote* sentences and R represents *Response* sentences in templates.

Category (number)	Indicators
Cue words (48)	so, oh, well, just, and, because, though, as well, if, then, thus, unless, seems, also, you, uh, still, seriously, what about, how about, in my view, in my opinion, after all, only, as long as, I think, I see, I know, in fact, so that, even, I mean, cause, at least, wonder, anyway, my perspective, we should, what if, should be, I'm afraid, for example, for instance, conclusion, summarize, consequence, for this reason, as a result
Agreement (145)	yes, correct, agree, accept, support, true, like, good, exactly, ok, okay, right, clear, sure, thanks, believe, of course, make sense, as well, favor, favoring, too, either, can, reasonable, exact, certainly, I believe, admit, clearly, make sure, prove, pretty, thank, definitely, consent, obvious, same, yeah, absolutely, convincing, clarify, point out, explain, enough, completely, acknowledge, claim, it seems to me, clean up, fully, help, reference, source, better, always, well said, accurate, consensus, explanation, acclaim, adequate, thoughtful, handy, pleased, sufficient, peaceful, enhance, appreciate, honor, friend, affirm ,trusted, stimulate, simple, hopeful, believable, nice, confirm, progress, fine, strong, rational, perfect, impress, suitable, decisive, motivate, interesting, respect, achieve, adorable, reaffirm, succeed, helpful, preferable, satisfied, confident, advantage, encourage, truthful, satisfy, steady, effective, success, benefit, useful, relevant, realistic, agreeable, ease, standout, generous, beneficial, capable, insight, doubtless, liking, grateful, outperform, reward, coherence, easy, improve, suffice, fans, consistent, relevance, supportive, logical, qualified, ideal, proper, glad, concise, best, proving, gain, approval, tidy, authentic, great, reliable, cool, faithful
Denial (278)	no, not, never, nothing, however, but, doesn't, don't, isn't, yet, none, hate, false, wrong, doubt, disagree, how can, I don't think, missing, how come, does not, do not, aren't, are not, is not, instead, ?, against, missed, unlikeness, didn't, did not, stop, disfavoring, why, neither, nor, can not, can't, harm, hurt, harmful, lie, worse, unreasonable, apologize, fault, nobody, I don't believe, deny, joke, kidding, or not, confuse, confusion, were not, weren't, hard, few, forget, misunderstanding, who knows, sorry, stupid, problem, dissent, refuse, incorrect, sad, not true, no evidence, do you mean, I don't know, I don't see, you don't know, we don't have, I am wondering, you don't understand, offend, nope, oppose, issue, while, limited, criticize, criticism, ridiculous, ludicrous, difficult, angry, damage, hedge, miss, clumsy, pity, kill, ineffective, offense, impatience, unbearable, strange, lack, afraid, complex, annoy, foolish, cons, lied, noisy, mess, disappoint, unfair, junk, stupidity, pretend, blurring, opponent, diss, inefficiency, frighten, broke, skeptical, hopeless, inequality, deject, restrict, concern, complain, controversial, inconsistence, disregard, mad, guilty, collapse, slow, degenerate, anxious, abnormal, unfaithful, contend, error, betraying, denial, badly, incomplete, disgust impossible, degrade, ashamed, accuse, accusation, lacking, bother, failure, bothering, corrupt, bias, troublesome, insane, absence, cheater, uncertain, burden, abusive, unavailable, defect, aggressive, dumb, concerned, deaf, poor, ignore, cheating, crack, unbelieve, confess, imbalance, sucked, unacceptable, conflict, impossibly, distrust, silly, pitiful, terrible, dislike, break, die, darken, unauthentic, dispute, odd, hating, puzzled, disbelieve, mistake, clueless, idiot, illogically, insignificant, uncomfortable, doubtful, ill, shit, bored, pain, contradiction, overwhelm, boring, inconsistent, delay, averse, allergic, stubborn, coarse, conflicted, anger, painful, arrogant, object, acerbic, ignorance, struggle, refused, disadvantage, confusing, complicated, disapproval, cheat, unconvincing, weak, stuck, annoying, objection, anti, inferior, contradict, mistrust, reject, drop, attack, adverse, liar, concede, rejected, fuck, disapprove, fail, ignorant, excuse, mistaken, cheated, abuse, afflict, unconfirmed, suck, aggravate, lacked, betray, idiotic, anxiety, loss, hateful, bait, punish, banish, illness, stupidly, ambiguous, idiocy, dangerous, skeptic, injury, illogic, illogical, blur, contradictory, trouble,
Hedge (25)	maybe, probably, would, could, rather, although, really, actually, wondering, possibly, essentially, anyway, somewhat, somehow, I suppose, perhaps, 'd like, would like, would rather, whatever, may, wouldn't, might, generally, personally

Table 11: Lexicon of domain-independent (dis)agreement indicators