Active Learning with Multi-annotator Disagreement

by

Mei Lan Schrama

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday August 26, 2022 at 14:00 PM.

Student number:	5280257	
Project duration:	November 1, 2021 – August 10, 2022	
Supervisors:	Prof. C. M. Jonker,	TU Delft, EEMCS
	Dr. P. K. Murukannaiah,	TU Delft, EEMCS
	Ir. E. Liscio,	TU Delft, EEMCS
Thesis committee:	Dr. J. Yang,	TU Delft, EEMCS

This thesis is confidential and cannot be made public until December 31, 2023.

TU Delft

Abstract

Moral values are often used as guidelines for human behaviour. The ability to identify moral values is important for social and ethical artificial intelligence. We address the difficulties of using contemporary natural language processing (NLP) techniques to classify moral values in texts. As the classification criterion of moral values is subjective, it is often difficult to argue for the existence of a 'ground truth' label. In such circumstances, we can learn from the (dis-)agreement among multiple annotators. However, it is expensive to consult everyone, especially when working with crowdsourcing. A way to reduce the annotation cost is to apply active learning, which uses query strategies to choose the data and annotator that are most valuable to consult. Further, to account for subjectivity, we want to ensure the dataset is labelled by a diverse set of annotators. Therefore, we propose an annotator selection method for active learning. When given an unlabelled text, this method selects an annotator that has labelled the least amount of texts that are similar to the given text. The evaluation results show that the method performs better on datasets with balanced annotator distribution.

1 Introduction

Artificial intelligence (AI) is augmenting human intelligence in a variety of domains such as manufacturing, transport, healthcare, and education. The actions AI systems take may have moral consequences for humans (Shank and DeSanti 2018; Wallach, Allen, and Smit 2008). Therefore, AI systems that perform morally relevant tasks should be equipped with the ability to understand human values. Values are beliefs and standards that guide actions, and they can explain the motivations for attitudes and behaviour (Schwartz 2012). By recognizing values, AI systems can align with human interests and goals (Russell, Dewey, and Tegmark 2015).

In recent years, the research intersection of ethics and natural language processing (NLP) is gaining attention. With the popularity of social media, people share their opinions and concerns about social problems on public platforms. This provides researchers with valuable opportunities to study the moral aspects from natural language. For example, existing studies (Araque, Gatti, and Kalimeri 2020; Hoover et al. 2020; Liscio et al. 2022; Alshomary et al. 2022) have contributed to text-based *value classification*, which identifies the moral values expressed in a piece of text. However, moral value classification is extremely challenging.

The first challenge of the value classification task comes from the limitation in the scale of data. Existing studies have shown that state-of-the-art neural language models are not capable of correctly inferring ethical norms from web text through self-supervision only (Jiang et al. 2021). It is necessary to have a labelled dataset as the moral textbook that can be used to teach machines to recognize values. Another feature of human values is that they can evolve over time. As society evolves, we should adjust and maintain the dataset to ensure the algorithm is up to date. Taking these problems into account, we can consider applying *active learning* (AL) for value classification tasks to reduce the human resources and costs in building such a dataset (Settles 2009). Compared to traditional machine learning, active learning is conducted in multiple iterations. In each iteration, it uses a query strategy to select a subset of most valuable data from an unlabeled data pool. By training the machine with selected data, active learning can yield better performance with less labelled data.

The second challenge is that the value annotations are subjective. The annotators may be influenced by different cultures and experiences, which give them different opinions towards values (Sap et al. 2022). This is a reason why annotators may disagree on the labels of the same piece of data. It is difficult to argue that there exist ground truths in subjective tasks (Uma et al. 2021). Furthermore, Aroyo and Welty (2015) emphasise that disagreement is not noise, but a signal that provide useful information for learning. Therefore, instead of finding a single correct answer in value classification task, it is also valuable to study the disagreement among multiple annotators. Although there exist methods of incorporating disagreement in the training of NLP models (Uma et al. 2021), no research has addressed actively seeking diverse annotations that the model can learn from.

We propose a method to actively seek diversity of annotators. In active learning, when two or more annotators are asked to annotate the same data point, they are usually selected either based on annotators' reliability or randomly. The former strategy is not suitable for a subjective task because we can not determine the reliability of an annotator when there is no ground truth label. However, if we select the annotators randomly, there is a risk of falling into the majority trap. We propose a method for actively avoiding that by smartly selecting annotators. When given an unlabelled text, this method selects an annotator that has labelled the least similar texts before. In this way, the NLP models can actively learn from a diverse set of annotators. The proposed method is tested on the MFTC dataset (Hoover et al. 2020). The evaluation results show that the method can outperform random annotator selection on datasets with balanced annotator distribution.

Organization Section 2 introduces the background knowledge and existing studies related to our topic. Section 3 describes the methods we apply. The setup of our experiment is described in Section 4, followed by results and discussion in Section 5. Section 6 draws conclusions and proposes future directions.

2 Literature Review

We introduce the related works on active learning, learning from disagreement and value classification.

2.1 Active Learning

Active learning is a subfield of machine learning in which the machine employs a *query strategy* to actively determine which data is the most valuable to be annotated next. Active learning requires less labelled data to achieve comparable performances as regular machine learning, thereby reducing the cost of time and labour. Active learning can be divided into membership query synthesis, stream-based selective sampling and pool-based sampling (Settles 2009). In the membership query scenario, the model generates queries to be labelled. Stream-based selective sampling inputs the data

sequentially and lets the model decide to query its label or not. For pool-based sampling, the model uses query strategies to sample a set of queries from a data pool. In this paper, we focus on pool-based active learning.

The query strategies of pool-based active learning can be categorized into three main categories: data-based, model-based and prediction-based strategies (Schröder and Niekler 2020). Data-based strategies involve only information about the input data, such as representativeness of the input space. This typically involves compressing data points into sets and using the least amount of representative data to achieve higher generalization (Sener and Savarese 2018; Nguyen and Smeulders 2004). Model-based strategies consider the information of both data and model. Well-known methods are model uncertainty (Gal, Islam, and Ghahramani 2017) and expected parameter change (Settles, Craven, and Ray 2007). These strategies use the model's parameters as measures of data selection, such as model weight and gradient. Finally, the prediction-based methods use the properties of predicted outputs as selection principles. In this case, the instances with the highest prediction uncertainty (Lewis and Gale 1994) or expected prediction change (Roy and McCallum 2001) are considered more valuable for querying. Apart from these three categories, random selection is commonly used as a baseline (Schröder and Niekler 2020).

For crowdsourced datasets, we can also involve annotator selection in the query strategy—that is, not only selecting the instance to be annotated, but also the annotator. In this case, the selection of data and annotator can be either sequential or joint. The sequential selection first selects the unlabelled data instances, and then selects the annotators to label them. For example, Rodrigues, Pereira, and Ribeiro (2014) treat the selection of instances and annotators separately. They model properties such as sensitivity and specificity of annotators, and use these measures to select the best annotator to label the query instances. In contrast, joint selection methods select instance and annotator in pair. For example, Yan et al. (2011) balance the trade-off between useful query and annotator accuracy to select a query-annotator pair. Yan et al. (2012) form a joint selection criterion by combining the query utility and annotator performance.

We are not aware of an annotator selection strategy that focuses on the annotators' disagreement. The related works are mostly evaluated under the assumption of the existence of a gold label. Thus, the existing query strategies for active learning aim to minimize the distance between a 'correct' ground truth label and the label provided by the classifier. However, for a value classification task, it is difficult to define ground truth labels because of the subjective classification criterion. Therefore, we want to consider the disagreement among multiple annotators.

2.2 Learning from Disagreement

In a crowdsourced dataset, one data point can be labelled by multiple annotators. However, the annotators may disagree, which means that different annotators provide different labels for the same data point. In this case, we need to find a way to deal with the disagreement. A simple approach is to use a *hard label* (also called a *gold label*) with majority aggregation. However, majority aggregation could be problematic because it would ignore the opinions of the minority. A recent survey (Uma et al. 2021) divided the existing methods to deal with disagreement into four categories: aggregation of coder judgements, filtering hard items, learning directly from crowd annotations, and augmenting hard labels with information from the crowd annotations. Widely used methods for aggregating annotations are majority voting (Dawid and Skene 1979), modelling the sensitivity, specificity or reliability of annotators (Carpenter 2008; Hovy et al. 2013), modelling the agreement or disagreement on instances (Inel et al. 2014; Dumitrache, Aroyo, and Welty 2018). Filtering hard items is to filter out the instances with low annotator agreement from the training and testing data, or separate them from high agreement data (Reidsma and op den Akker 2008; Klebanov, Beigman, and Diermeier 2008). These two categories only involve hard labelling, which assumes that there exists a binary ground truth, i.e., each data point is labelled with either 1 (true) or 0 (false).

To learn directly from crowd annotations, a *soft label* (Peterson et al. 2019; Uma et al. 2020) is widely used. Compared to hard labels, soft labels can use fractional numbers to represent the annotations. It can be used to model the distribution of the annotations. In this project, we use the probability distribution of all annotators' annotations as the soft label. For example, consider the case in which two annotators out of three annotated 1 (true) for a data point, while the other annotator provided a 0 (false) label. We integrate these annotations into one soft label of 0.66. Another way to learn directly from crowd annotations is repeated labelling (Sheng, Provost, and Ipeirotis 2008). It inputs the same data point multiple times, each time with the labelling from a different annotator. Finally, there are also methods that make use of both hard labels and information about disagreement. For example, weighing the loss associated with an instance by an estimate of the uncertainty on its labels (Plank, Hovy, and Søgaard 2014), or training the model with both gold labels and soft labels (Lalor, Wu, and Yu 2018).

Several metrics have been proposed to evaluate learning from disagreement. Hard evaluation metrics, such as accuracy or F1-scores, assume the existence of binary ground truths. Soft evaluation metrics do not assume binary ground truths. For example, soft evaluation can measure the similarity between the actual annotation and classifier's prediction via metrics such as cross-entropy and Jensen-Shannon divergence (Lin 1991) between the probability distribution of annotations from multiple annotators and the machine prediction (Peterson et al. 2019). Uma et al. (2021) use entropy similarity and entropy correlation to measure the ability of the model to capture human disagreement in its prediction.

Based on our literature review, we identify a research gap in active learning with multi-annotator disagreement. Exploring this research gap could help improve the application of active learning on tasks involving subjective judgements, such as hate speech detection (Akhtar, Basile, and Patti 2019) and recognizing ethical norms (Jiang et al. 2021). Moreover, active learning could reduce the scale of labelled datasets, which is also saving the costs of time and labour we need. Therefore, we propose a method for seeking annotator

diversity in active learning, and use the described methods to learn from the annotators' disagreement.

2.3 Value Classification from Text

Text-based value classification has been studied with both unsupervised and supervised machine learning. Unsupervised methods (Rezapour, Shah, and Diesner 2019; Araque, Gatti, and Kalimeri 2020; Hopp et al. 2021) use value lexicons to identify moral values. The supervised machine learning models (Lin et al. 2018; Hoover et al. 2020; Liscio et al. 2022) are trained with textual datasets labelled with moral annotations.

Most studies use majority vote to aggregate labels from annotators when encountering disagreement (Hoover et al. 2020; Araque, Gatti, and Kalimeri 2020; Liscio et al. 2022). Araque, Gatti, and Kalimeri (2020) filter out the annotators that have low agreement with other annotators and high inconsistent ratings, and then aggregate the annotations.

Active learning is also based on supervised machine learning, but it also requires unlabelled data to query. Existing research has not yet applied active learning in value classification tasks. We apply a proposed active learning method in text-based value classification tasks to learn from the disagreement among annotators.

3 Methodology

We describe the active learning framework, the proposed method and query strategies used.

3.1 Active Learning Framework

Our method is based on an active learning framework. Given a labelled training dataset X and an unlabelled dataset (or data pool) U, active learning employs the following training process:

1. Train the model with the labelled dataset X.
2. In the unlabeled dataset U, apply a query strategy to select the data that we want to use and the annotators to label them. A subset S is selected from U. Ask the selected annotators to label dataset S.
3. Add the data and labels of S into dataset X, and remove S from U: $X = X + S, U = U - S$.
4. Return to the first step until the model reaches the stop criterion.

After the training iterations, the model is evaluated on a labelled testing dataset Y. For active learning, we can adjust the query strategy to change the data and annotations that are used for training the model.

3.2 Proposed method

We propose a sequential query strategy to select the train data and annotators. In this case, the query strategy can be divided into two parts: data selection and annotator selection. We use an popular strategy named uncertainty sampling for data selection (Lewis and Gale 1994). The reason why we choose uncertainty sampling and how it works are introduced in the next section.

Our main contribution is the annotator selection. Since we aim to represent the opinions of all of our annotators, we want to ensure that each annotator can label a diverse set of data. Therefore, we propose a new annotator selection strategy named DIVA (DIVerse Annotator selection). When a new data point is selected by the data selection strategy, DIVA ensures annotator diversity by selecting an annotator who has not labelled similar data before. DIVA uses a separate NLP model to select the annotator, different from the model used for value classification. We explain this further in Section 3.3.

Figure 1 shows how we apply the query strategies and train the models. The green blocks indicate the data selection processes. The yellow blocks show annotator selection.

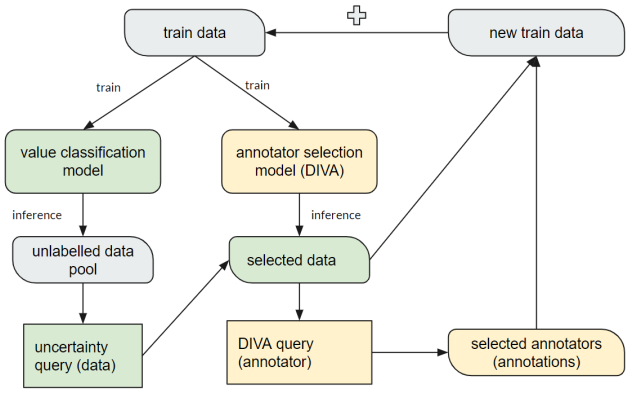


Figure 1: Overview of the proposed method

The proposed method works as follows:

1. Train the value classification model and DIVA model with the labelled data.
2. Conduct inference on unlabelled data pool with value classification model and apply uncertainty query to select the new subset to be labelled.
3. Conduct inference on the selected subset with DIVA model and apply DIVA query to select an annotator for each data point. Get the annotations from the selected annotators.
4. Add the selected data and annotations to the labelled data.
5. Return to the first step until the model reaches the stop criterion.

3.3 Query strategy

We explain the strategies for data and annotator selection.

Data selection We use uncertainty sampling (Lewis and Gale 1994) to select the text data. It is a widely used prediction-based strategy. In our project, we use the value classification model to predict the moral labels for texts in the unlabelled data pool. Then we can calculate how certain these predictions are using the following equations.

$$U(x_c) = \begin{cases} 1 - P(x_c), & P(x_c) > 0.5, \\ P(x_c), & P(x_c) \leq 0.5 \end{cases} \quad (1)$$

$$U(x) = \sum_{c \in C} U(x_c) \quad (2)$$

In these equations: C represents the moral label classes, $P(x_c)$ is the probability prediction of class c for text x, and $U(x)$ is the uncertainty level of text x. We select a subset of the unlabelled data pool with higher uncertainty levels.

We choose uncertainty sampling as our data selection strategy based on its actual performance on our task. We compared its performance with another state-of-art query strategy named core-set selection (Sener and Savarese 2018). Uncertainty sampling outperformed both core-set selection and baseline random selection. More details can be found in Appendix A.1.

Annotator selection We propose a new query strategy DIVA for diverse annotator selection. DIVA models the relationship between text similarity and annotator. When a new text is selected, DIVA uses this relationship to select an annotator that has labelled the least similar texts before. Similar data should not be labelled by the same annotator too many times because this will make the model biased. The model will learn too much from that single annotator and ignore the opinions of other annotators. Therefore, we propose this strategy to ensure that the training data is labelled by a diverse set of annotators. In this way, we can reduce the bias and learn from the diversity in annotations.

Figure 2 shows how DIVA model is used in active learning. It is trained with a text input and corresponding *annotator label*. The annotator label indicates who has provided value annotation for that text data. For example, for the input of initial training in Figure 2, $text1$ is labelled by only annotator c. The binary label of annotator c corresponding to $text1$ is 1, while other annotator labels are 0. Thus, the full annotator label of $text1$ is [0 0 1 0]. We treat these as the training data and labels of the DIVA model. Our goal is to have DIVA learn that we already considered the annotations of annotator c for texts that are similar to $text1$.

At the beginning of active learning, we train the DIVA model with randomly selected text data and annotator labels. In this way, the model can learn the initial distribution of annotators and texts. After training, we move on to a new iteration in active learning. We conduct inference on a selected unlabelled dataset with the trained DIVA model. As shown in the figure, the inference result indicates that annotator d scores the least for $text51$. This shows that annotator d has labelled the least amount of texts that are similar to $text51$. In contrast, annotator c has the highest score of 0.4, which means that in the input data of the earlier training, data points that are similar to $text51$ have been mostly labelled by annotator c. If we would ask annotator c to label $text51$, the model would become more and more biased. Therefore, we select annotator d to provide annotations for this text. Similarly, we select annotator a for $text52$. Therefore, the annotator labels for $text51$ and $text52$ are [0 0 0 1] and [1 0 0 0] respectively. We combine these data points with

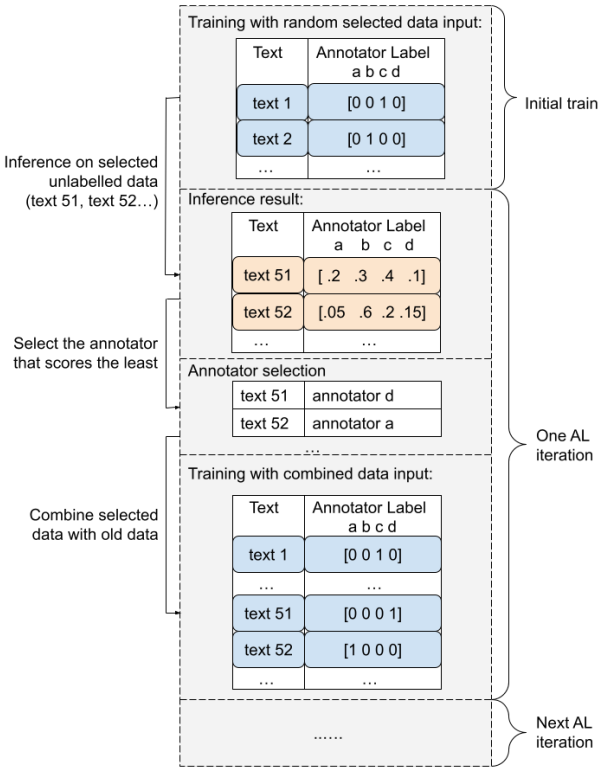


Figure 2: Overview of DIVA (DIVerse Annotator selection)

the training input from the previous iteration and generate a new labelled dataset. At the end of this active learning iteration, we train the DIVA model with this combined dataset. After this, the DIVA model will be used to infer other new data in the next active learning iteration.

4 Experiment

We introduce the dataset used for experiments. We explain the experiment settings, evaluation metrics and hypotheses.

4.1 Dataset

Our experiment uses the Moral Foundations Twitter Corpus (MFTC) (Hoover et al. 2020). MFTC measures moral sentiment in natural language. They use the moral categories from Moral Foundations Theory (MFT) (Graham et al. 2013) as the value labels of the dataset. The MFT is a social psychological theory that proposes five primary foundations of moral concerns to represent people's moral values. Each of the five foundations is composed of a vice and virtue label. Table 1 describes the MFT foundations.

MFTC is composed of 35108 labelled tweets. The tweets come from seven different domains, each refers to a different topic: All Lives Matter (ALM), Baltimore protests (Baltimore), Black Lives Matter (BLM), hate speech and offensive language (Davidson), 2016 presidential election (Election), MeToo movement (MeToo) and Hurricane Sandy (Sandy). Each tweet is labelled by at least three and at most eight annotators. In total, the dataset is labelled by 23 different

Table 1: The five moral foundations in the MFT

Foundation	Definition
Care/ Harm	Encourage caring for others/ Avoid harming others
Fairness/ Cheating	Encourage defending fairness and equality/ Avoid cheating or exploiting others
Loyalty/ Betrayal	Encourage prioritizing one's in-group/ Avoid betraying or abandoning one's in-group
Authority/ Subversion	Encourage obeying authority and tradition/ Avoid subverting authority or tradition
Purity/ Degradation	Encourage the purity of sacred entities/ Avoid contamination of such entities

annotators. Each tweet can be labelled with multiple values. Also, the MFTC uses a *Nonmoral* label to express that the data is not related to any moral labels in MFT.

Table 2 shows how many tweets are labelled by different numbers of annotators. Table 3 shows the number of annotations provided by each annotator for each MFTC domain. More detailed distributions are included in Appendix A.2.

4.2 Experimental setup

We use BERT as the NLP model for both the value classifier and DIVA method. It is a widely used state-of-art model in NLP tasks (Devlin et al. 2018). To evaluate the effectiveness of the proposed method, we compare its performance with that of a baseline strategy. Both the baseline strategy and our proposed method use uncertainty sampling to select text data. The difference is that baseline strategy uses randomly selected annotators to label the text data, while the proposed method uses DIVA query to select annotators. Since both the uncertainty sampling and DIVA query require the inference results from trained models to start, we initialize the training with random text data and annotators, as it is common to use randomly selected data for initial training in general active learning tasks (Yuan et al. 2011). During experiments, it was found that for different domains, the model requires different amounts of initial data to generate valid results. During testing with different data-size settings, we find that if the initial training set has less than 300 texts, the f1 scores will be insignificant for most datasets. Therefore, we decide to conduct initial training with 300 data and select 100 new data for each iteration of active learning. In this way, we can see the differences in the performance from a small amount of data to relatively larger data size, and how it changes every step. For some text data, the MFTC dataset only collected annotations from 3 annotators. Therefore, we select 1 annotator to provide annotation for each unlabelled text to avoid empty queries and enhance the difference between DIVA query and random selection. The models are trained for 10 active learning iterations. We apply 5-fold cross-validation to evaluate the results. Then the models are evaluated with the metrics explained in Section 4.3.

4.3 Evaluation

We employ the following evaluation metrics.

Hard evaluation Hard evaluation assumes there exists a ground truth even though there is disagreement among annotators. In this project, we aggregate true labels with the majority vote and use F1 scores to measure the model's ability to predict the majority agreement. F1 score is widely used to measure the accuracy of multi-label classification tasks. It is calculated from the precision and recall of the prediction compared to the true label. For our task, the distribution of label classes is not equally balanced, so we use both micro-F1 and macro-F1 scores. Macro-F1 computes the precision and recall for each class independently, and then takes the average. In contrast, micro-F1 computes the average result for all the classes together. Therefore, micro-F1 score is more dependent on the major class, while macro-F1 treats all classes equally.

Soft evaluation To measure the diversity of the model's prediction, we cannot drop the minority opinions. Therefore, we also use soft evaluation, which does not assume the existence of a correct label. Peterson et al. (2019) proposed using cross-entropy to measure the similarity between the distributions of the model's prediction and the actual annotations from multiple annotators. The cross-entropy of two distributions is calculated as:

$$H(p, q) = -\sum p(x) \times logq(x), \tag{3}$$

where $p(x)$ is the probability distribution of the annotations, and $q(x)$ is the predicted probability distribution. Cross entropy measures how similar these two distributions are. For our task, we represent the diversity of annotations from multiple annotators with a soft label generated from the probability distribution of actual annotations. In this way, we measure the model's ability to learn this diversity as the cross-entropy loss between its prediction and the actual soft labels.

To study a subjective task, we need to consider not only the majority agreement, but also the opinions of other annotators. Since we use the majority vote to generate gold labels to calculate F1 scores, the results of hard evaluation can reflect how well the model predicts the majority agreement. In contrast, soft evaluation does not ignore the opinions of the minority. It can measure the model's ability to predict the annotations from all annotators. If the models are evaluated only using either hard or soft evaluation metrics, we will only get a partial view of how they perform. (Uma et al. 2021) Therefore, we apply both hard evaluation and soft evaluation to measure the efficiency of the proposed method.

4.4 Hypotheses

The following are our hypotheses.

Hypothesis 1 We expect the proposed method to have a more significant improvement for soft evaluation than hard evaluation. This is because ensuring each annotator labels a diverse dataset should help the model learn from all annotators in balance. Therefore, it is more beneficial for soft eval-

Table 2: Counts of tweets annotated by different numbers of annotators.

#Annotators	ALM	Baltimore	BLM	Davidson	Election	MeToo	Sandy
3	4316	4496	28	4959	659	2522	4591
4	108	575	388	2	4699	2006	--
5	--	522	4837	--	--	62	--
6	--	--	--	--	--	295	--
7	--	--	--	--	--	5	--
8	--	--	--	--	--	1	--

Table 3: Number of annotations provided by different annotator over corpus domains

Annotator id	ALM	Baltimore	BLM	Davidson	Election	MeToo	Sandy	Total
00	113	0	5067	0	5097	0	0	10277
01	4225	86	5152	73	51	2042	158	11787
02	4384	4959	5180	0	5033	0	0	19556
03	4410	0	5239	0	5319	0	0	14968
04	1	0	5199	0	5264	18	0	10482
05	0	0	0	4662	0	7	0	4669
06	0	0	0	4089	0	0	0	4089
07	0	0	0	4738	0	0	0	4738
08	0	0	0	4805	0	0	0	4805
09	0	0	0	0	0	0	4591	4591
10	0	0	0	0	0	0	4588	4588
11	0	0	0	0	0	0	4555	4555
12	0	1228	0	0	0	0	0	1228
13	0	5420	0	0	0	0	0	5420
14	0	5416	0	0	0	0	0	5416
15	0	673	0	0	0	0	0	673
16	0	0	0	0	0	2499	0	2499
17	247	616	8	69	9	2764	17	3730
18	0	0	0	0	0	2592	0	2592
19	0	0	0	0	0	2478	0	2478
20	0	0	0	0	0	2591	0	2591
21	0	0	0	0	0	2162	0	2162
22	0	0	0	0	0	560	0	560

uation as it considers the labels from all annotators, while hard evaluation only considers the majority agreement.

Hypothesis 2 Table 2 indicates that BLM has 4837 tweets that are labelled by 5 annotators, and this is the most among all datasets. This means for these tweets, random selection and DIVA strategy are more likely to select different annotators. Similarly, compared to other datasets, MeToo and Election datasets have the most amount of tweets that are labelled by 4 annotators. Therefore, we expect the difference between the performances of these two strategies to be more significant in BLM, MeToo and Election datasets.

Hypothesis 3 In Table 3, we can see that the distributions of annotations are also different for each domain. Among all domains, BLM has the most balanced distribution. For some domains, the differences between the highest and lowest numbers of annotations are so large that it is reasonable to assume that some annotators could not influence the result. For example, in MeToo domain, $annotator04$ labelled 18 tweets, while $annotator22$ and $annotator17$ labelled 560 and 2764 tweets respectively. Then $annotator04$ may not influence the final prediction because the model does not have enough train data for $annotator04$. In contrast, $annotator22$ and $annotator17$ can make more significant differences in the results. Based on this observation, the BLM and Election domains have more balanced annotator distributions compared to other datasets. Since we split the training and testing sets randomly, they are likely to have a similar annotator distribution. This means if we have a biased training set where a certain annotator has labelled a lot more data than other annotators, the influence of this annotator will also be dominant in the testing set. In this case, the baseline strategy (random selection) will select a similarly biased subset from the training data, while the proposed method will try to avoid the bias and select diversely. However, since the testing set is unbalanced, training the model with a diversely selected dataset may overemphasise the annotators that labelled fewer tweets. This will decrease the similarity between the annotations of the testing set and the prediction of the classifier. Therefore, we expect the proposed method to achieve better results on balanced datasets for both hard and soft evaluations.

5 Results and Discussion

We conducted experiments on all 7 domains of MFTC. Parts of the evaluation results are plotted in Figures 3 and 4. These figures show the evaluation of how the methods perform on the testing set at different iterations of active learning. The x-axis indicates the data size of the training set. Each different data size represent a different active learning iteration. The y-axis shows the performance of the two models.

Contrary to our first hypothesis, the differences between these two strategies are more significant when using hard evaluation (micro and macro F1-scores) compared to soft evaluation (cross-entropy). For the results of the BLM domain, we can see that in Figure 3a, the difference between micro-F1 scores for the two methods is significant at datasize 500, while in Figure 3c it is not significant. Similar observations can be found in Figure 4d and 4f at datasize 800. We believe that this is because the cross entropy metric we use for soft evaluation is not the most suitable metric for this project. It might be not capable enough to measure the difference in diversity between the two methods. After all, there is not a metric that is generally accepted as optimal for soft evaluation yet (Uma et al. 2021). However, soft evaluation can still provide an estimation of how close the model's prediction is to the actual annotations.

As we expected in hypothesis 2, the differences between the baseline strategy and proposed strategy are not significant for Baltimore, Davidson and Sandy datasets. This is because there are fewer annotators for each tweet in these datasets. Therefore, we only discuss the results of ALM, BLM, Election and MeToo datasets. The results for other datasets can be found in Appendix A.3.

Aligning to hypothesis 3, the DIVA strategy outperforms the baseline strategy for BLM and Election domain in Figure 3. In Figure 4, it is also observed that the proposed method does not work as well as the baseline strategy for ALM and MeToo datasets. According to our hypothesis, it could be that ALM and MeToo domains have imbalanced annotator distributions. However, the inter-annotator agreement are also different for these domains. As shown in Table 4, BLM and Election have higher inter-annotator agreement than ALM and MeToo. Although the proposed method does not consider how each annotator labels the data, the inter-annotator agreement level may be an influencing factor of this result. The agreement level can indicate how difficult the task is. For MeToo and ALM datasets, the graphs are not converged at the end of the training, which means the task is too difficult to learn with this amount of data. In this case, enhancing the diversity could bring negative impacts to the convergence. For active learning, how to achieve the desired result with a smaller datasize is also an important evaluation indicator. Therefore, we did not continue the training with more data but draw a conclusion that the proposed method does not perform well on these datasets.

5.1 Limitations

We discuss three limitations of our experiment. Firstly, we conducted the experiment only on MFTC. MFTC is composed of seven domains. We compared the performance of our methods in different domains and drew conclusions based on the differences between these domains. For example, the inter-annotator agreement, annotator number and annotator distribution can be the dependent variables that affect the experiment results. However, it is difficult to determine which domain properties caused the difference in performance. Therefore, the results may be dependent on these specific datasets, and they may not generalize to other datasets.

Table 4: Interannotator Agreement: Kappa (Fliess' kappa) and prevalence- and bias-adjusted kappa (PABAK)

	BLM	Election	ALM	MeToo
Kappa	0.38	0.29	0.27	0.21
PABAK	0.41	0.40	0.29	0.23

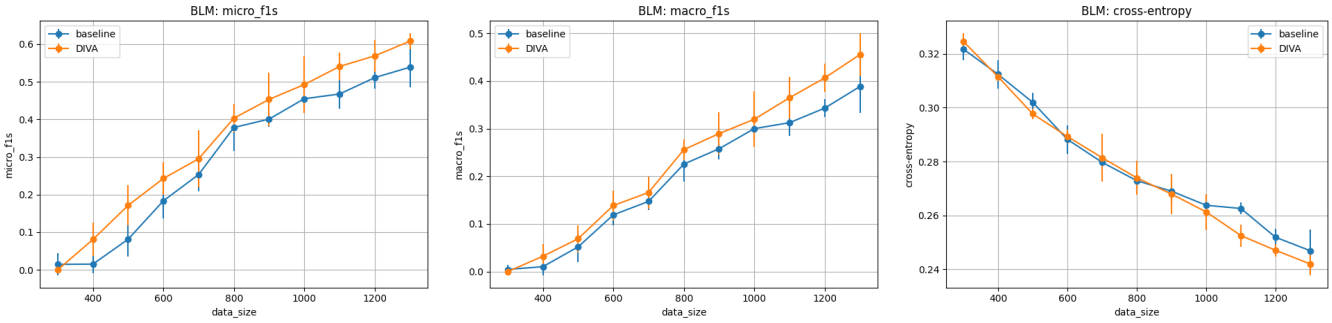
Secondly, the MFTC datasets have low inter-annotator agreement. It is also interesting to evaluate the method on datasets with a higher agreement level. We also only had data of at most 5 annotators per tweet. As explained in section 4.4, the number of annotations for each tweet influences how significant the results are. Therefore, it is necessary to conduct experiments on datasets with higher agreement levels and annotator numbers.

Finally, we compared our method with only a random baseline strategy. Since the existing annotator selection strategies do not focus on disagreement among annotators, their abilities to learn from disagreement are also yet to be evaluated. To make the experiment more convincing, we should involve the state-of-art annotator selection strategies as comparison objects in the future.

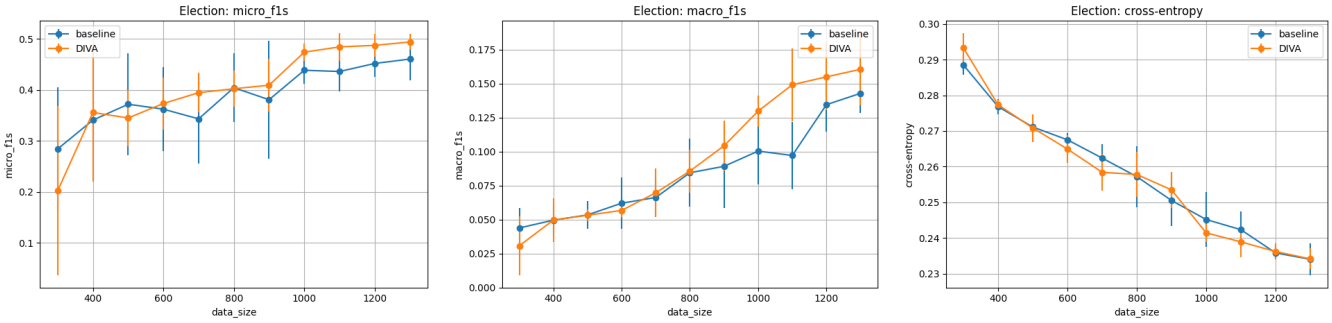
6 Conclusion and Future Work

The goal of text-based moral value classification is to identify which moral aspect is expressed in a piece of text. Using value classification, we can train AI to align with human values and enhance its ability to cooperate with humans. The challenges of moral value classification tasks lie in the subjective classification criterion and limitation of data scale. To address these problems, we try to learn from the annotators' disagreement with active learning.

We propose a new annotator selection strategy (DIVA) that aims to increase the diversity of annotators in active learning. We evaluated the proposed method on the MFTC dataset and compared its performance with random annotator selection as a baseline. The improvements are significant for datasets that have enough tweets labelled by more than 3 annotators. Furthermore, the metrics of hard evaluation (F1-scores) provide more significant differences than soft evaluation (cross-entropy). We observe that the proposed strategy has inconsistent performance on different datasets. We discussed two possible reasons. Firstly, the datasets have different annotator distributions. It is expected that the DIVA method performs better on the datasets that have more balanced distributions. Secondly, the task is more difficult for some datasets because they have lower inter-annotator agreement. In this case, the diversity DIVA brings

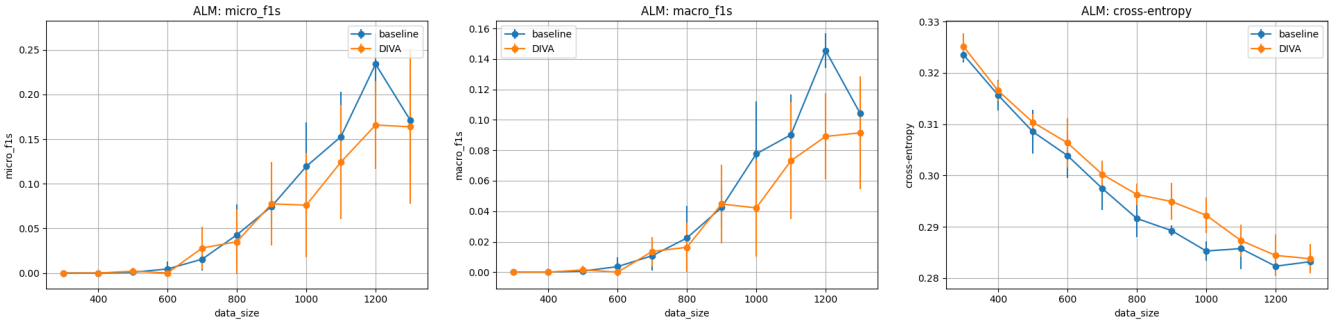


(a) Micro-f1 score (BLM) (b) Macro-f1 score (BLM) (c) Cross-entropy loss (BLM)

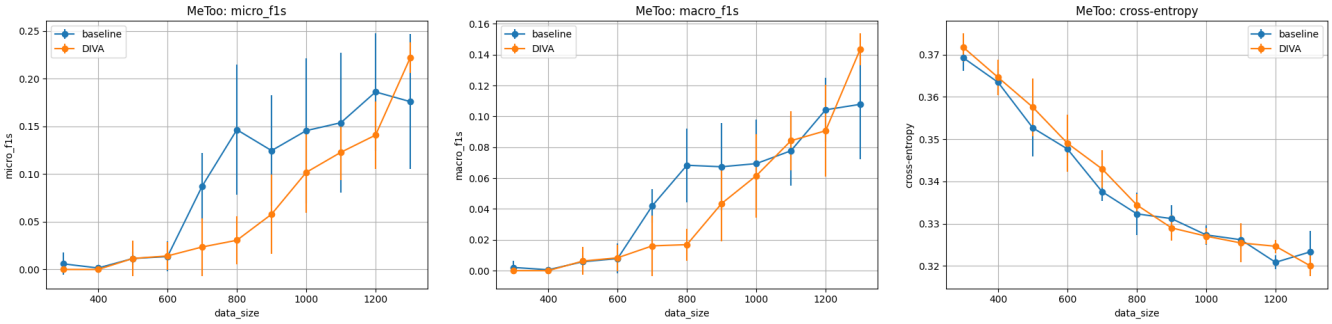


(d) Micro-f1 score (Election) (e) Macro-f1 score (Election) (f) Cross-entropy loss (Election)

Figure 3: Results for BLM and Election domains



(a) Micro-f1 score (ALM) (b) Macro-f1 score (ALM) (c) Cross-entropy loss (ALM)



(d) Micro-f1 score (MeToo) (e) Macro-f1 score (MeToo) (f) Cross-entropy loss (MeToo)

Figure 4: Results for ALM and MeToo domain

to the model may slow down the convergence.

We address a research gap in learning multi-annotator diversity in active learning tasks. Exploring this field can support the application of subjective machine learning tasks with limited labelled data. A future direction is to conduct experiments on other datasets to better determine which properties are influencing the results. This could bring new directions for improvement. Then, we could try other data selection strategies since data selection is an important part of the active learning framework. It would be also interesting to use joint selection instead of sequential selection. Selecting the data and annotator in pairs at the same time means we can consider the diversity of annotators one phase earlier. Furthermore, if we want to apply the model in practice, it is also important to adapt it to different degrees of inter-annotator agreement. Future studies could additionally take how the annotators label the data into account when selecting them. Finally, we could also see that cross-entropy we used for soft evaluation could not measure the difference between the two methods as well as F1 scores. Therefore, another research direction is to study new soft evaluation metrics for subjective machine learning tasks.

References

Akhtar, S.; Basile, V.; and Patti, V. 2019. A New Measure of Polarization in the Annotation of Hate Speech. In Alviano, M.; Greco, G.; and Scarcello, F., eds., *AI*IA 2019 – Advances in Artificial Intelligence*, 588–603. Cham: Springer International Publishing. ISBN 978-3-030-35166-3.

Alshomary, M.; El Baff, R.; Gurcke, T.; and Wachsmuth, H. 2022. The Moral Debater: A Study on the Computational Generation of Morally Framed Arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8782–8797. Dublin, Ireland: Association for Computational Linguistics.

Araque, O.; Gatti, L.; and Kalimeri, K. 2020. MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, 191: 105184.

Aroyo, L.; and Welty, C. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1): 15–24.

Carpenter, B. 2008. Multilevel bayesian models of categorical data annotation. *Unpublished manuscript*, 17(122): 45–50.

Dawid, A. P.; and Skene, A. M. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1): 20–28.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Dor, L. E.; Halfon, A.; Gera, A.; Shnarch, E.; Dankin, L.; Choshen, L.; Danilevsky, M.; Aharonov, R.; Katz, Y.; and Slonim, N. 2020. Active learning for BERT: an empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7949–7962.

Dumitrache, A.; Aroyo, L.; and Welty, C. 2018. Crowdsourcing ground truth for medical relation extraction. *ACM Transactions on Interactive Intelligent Systems*, 8(2): 1–20.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, 1183–1192. PMLR.

Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, 55–130. Elsevier.

Hoover, J.; Portillo-Wightman, G.; Yeh, L.; Havaldar, S.; Davani, A. M.; Lin, Y.; Kennedy, B.; Atari, M.; Kamel, Z.; Mendlen, M.; et al. 2020. Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8): 1057–1071.

Hopp, F. R.; Fisher, J. T.; Cornell, D.; Huskey, R.; and Weber, R. 2021. The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior research methods*, 53(1): 232–246.

Hovy, D.; Berg-Kirkpatrick, T.; Vaswani, A.; and Hovy, E. 2013. Learning Whom to Trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1120–1130. Atlanta, Georgia: Association for Computational Linguistics.

Inel, O.; Khamkham, K.; Cristea, T.; Dumitrache, A.; Rutjes, A.; van der Ploeg, J.; Romaszko, L.; Aroyo, L.; and Sips, R.-J. 2014. Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In *International semantic web conference*, 486–504. Springer.

Jiang, L.; Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Forbes, M.; Borchardt, J.; Liang, J.; Etzioni, O.; Sap, M.; and Choi, Y. 2021. Delphi: Towards machine ethics and norms. *CoRR*, abs/2110.07574.

Klebanov, B. B.; Beigman, E.; and Diermeier, D. 2008. Analyzing disagreements. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, 2–7.

Lalor, J. P.; Wu, H.; and Yu, H. 2018. Soft label memorization-generalization for natural language inference. *Proc. of UAI UDL Workshop*.

Lewis, D. D.; and Gale, W. A. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*, 3–12. Springer.

Lin, J. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1): 145–151.

Lin, Y.; Hoover, J.; Portillo-Wightman, G.; Park, C.; Dehghani, M.; and Ji, H. 2018. Acquiring background knowledge to improve moral value prediction. In *2018 ieee/acm international conference on advances in social networks analysis and mining (asonam)*, 552–559. IEEE.

Liscio, E.; Dondera, A.; Geadau, A.; Jonker, C.; and Murukannaiah, P. 2022. Cross-Domain Classification of Moral Values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 2727–2745.

Nguyen, H. T.; and Smeulders, A. 2004. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, 79.

Peterson, J. C.; Battleday, R. M.; Griffiths, T. L.; and Russakovsky, O. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9617–9626.

Plank, B.; Hovy, D.; and Søgaard, A. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 742–751.

Reidsma, D.; and op den Akker, R. 2008. Exploiting 'subjective'annotations. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, 8–16.

Rezapour, R.; Shah, S. H.; and Diesner, J. 2019. Enhancing the measurement of social effects by capturing morality. In *Proceedings of the tenth workshop on computational approaches to subjectivity, sentiment and social media analysis*, 35–45.

Rodrigues, F.; Pereira, F.; and Ribeiro, B. 2014. Gaussian process classification and active learning with multiple annotators. In *International conference on machine learning*, 433–441. PMLR.

Roy, N.; and McCallum, A. 2001. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, 2: 441–448.

Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, 36(4): 105–114.

Sap, M.; Swayamdipta, S.; Vianna, L.; Zhou, X.; Choi, Y.; and Smith, N. A. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5884–5906. Seattle, United States: Association for Computational Linguistics.

Schröder, C.; and Niekler, A. 2020. A survey of active learning for text classification using deep neural networks. *CoRR*, abs/2008.07267.

Schwartz, S. H. 2012. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1): 2307–0919.

Sener, O.; and Savarese, S. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*.

Settles, B. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison. Computer Sciences Technical Report 1648.

Settles, B.; Craven, M.; and Ray, S. 2007. Multiple-instance active learning. *Advances in neural information processing systems*, 20: 1289–1296.

Shank, D. B.; and DeSanti, A. 2018. Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in human behavior*, 86: 401–411.

Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 614–622.

Uma, A.; Fornaciari, T.; Hovy, D.; Paun, S.; Plank, B.; and Poesio, M. 2020. A case for soft loss functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 173–177.

Uma, A. N.; Fornaciari, T.; Hovy, D.; Paun, S.; Plank, B.; and Poesio, M. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72: 1385–1470.

Wallach, W.; Allen, C.; and Smit, I. 2008. Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *Ai & Society*, 22(4): 565–582.

Yan, Y.; Rosales, R.; Fung, G.; and Dy, J. G. 2011. Active learning from crowds. In *ICML*.

Yan, Y.; Rosales, R.; Fung, G.; Farooq, F.; Rao, B.; and Dy, J. 2012. Active learning from multiple knowledge sources. In *Artificial Intelligence and Statistics*, 1350–1357. PMLR.

Yuan, W.; Han, Y.; Guan, D.; Lee, S.; and Lee, Y.-K. 2011. Initial training data selection for active learning. In *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*, 1–7.

Zhang, L.; and Zhang, L. 2019. An ensemble deep active learning method for intent classification. In *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence*, 107–111.

A Appendix

A.1 Pre-test for data query strategy

We tried a state-of-art strategy named core-set selection (Sener and Savarese 2018). It samples a subset of unlabelled data that can best represent the whole dataset. The core-set selection problem is equivalent to a k-Center problem, which can be defined as choosing a set of centre points that minimise the largest distance between a data point and its nearest centre point. Sener and Savarese (2018) proposed a greedy algorithm to solve the problem. Existing works (Dor et al. 2020; Zhang and Zhang 2019) have applied the core-set selection method on NLP classification tasks using BERT. The results of both works showed the method achieved better performance than the random baseline model and was competitive among the state-of-art query strategies. Therefore, we thought it would be suitable for our task as well. Before applying it to our final experiment, we did a pre-test on the same task, but without involving our proposed annotator selection strategy. In the pre-test, we compared the results for three data selection strategies: random selection, uncertainty selection and core-set selection. The experiment settings are similar to our final test described in section 4.2. The difference is that we use only random annotator selection for this pre-test. Because of time constraints, we decide to conduct the pre-test only on the BLM dataset. We choose BLM because it has the most balanced distribution of value labels and annotators.

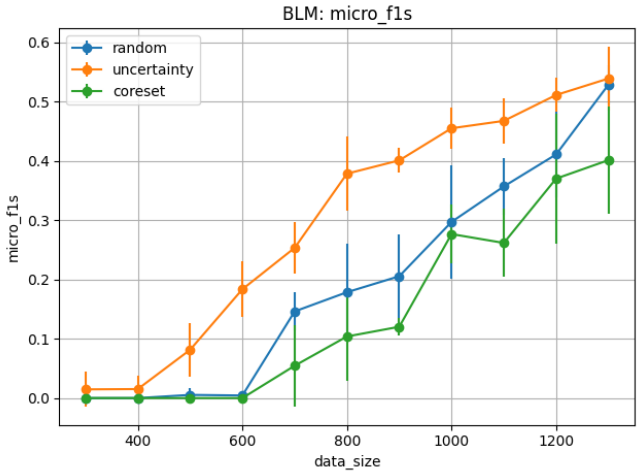
As shown in Figure 5, its actual performance is not comparable to random selection. In contrast, uncertainty sampling outperformed both the core-set and random selection in F1 scores and cross-entropy at most of the active learning iterations. Because our main focus is still on annotator selection, we could not try other state-of-the-art data selection strategies due to time limitations. Therefore, we decide to continue our project with uncertainty sampling.

A.2 Value distribution

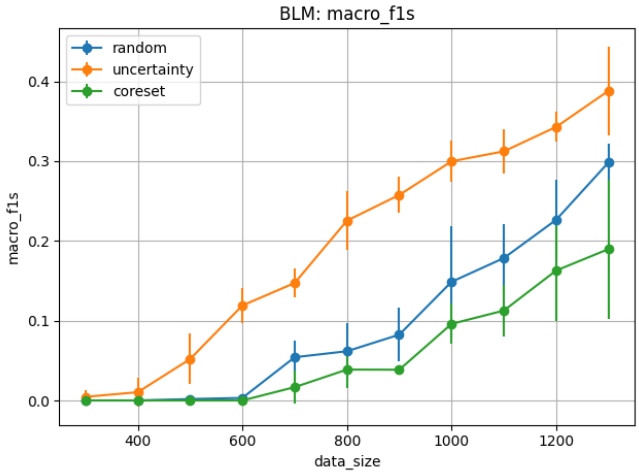
Table 5 indicate that how many value labels are provided for each domain. As shown in the table, the BLM domain has the most balanced distribution of labels, while the Baltimore and Davidson domains are highly unbalanced. Therefore, we applied both micro and macro F1 scores to measure the results.

A.3 Experiment results

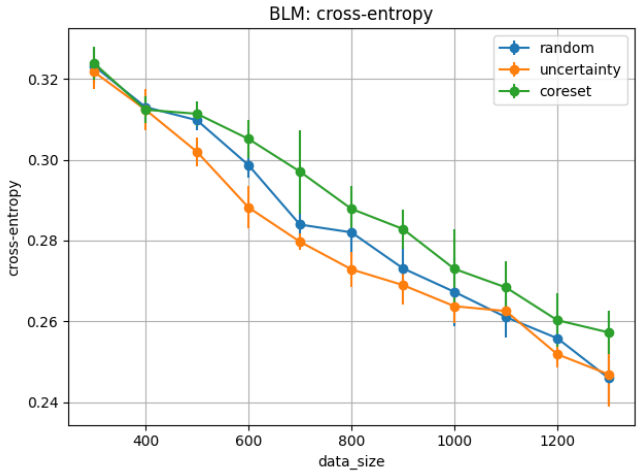
Figure 6 shows the evaluation results for Baltimore, Davidson and Sandy domains. For these datasets, the difference between our proposed method and baseline strategy is not significant. We discuss the reason in section 4.4.



(a) Micro-f1 score (BLM)



(b) Macro-f1 score (BLM)

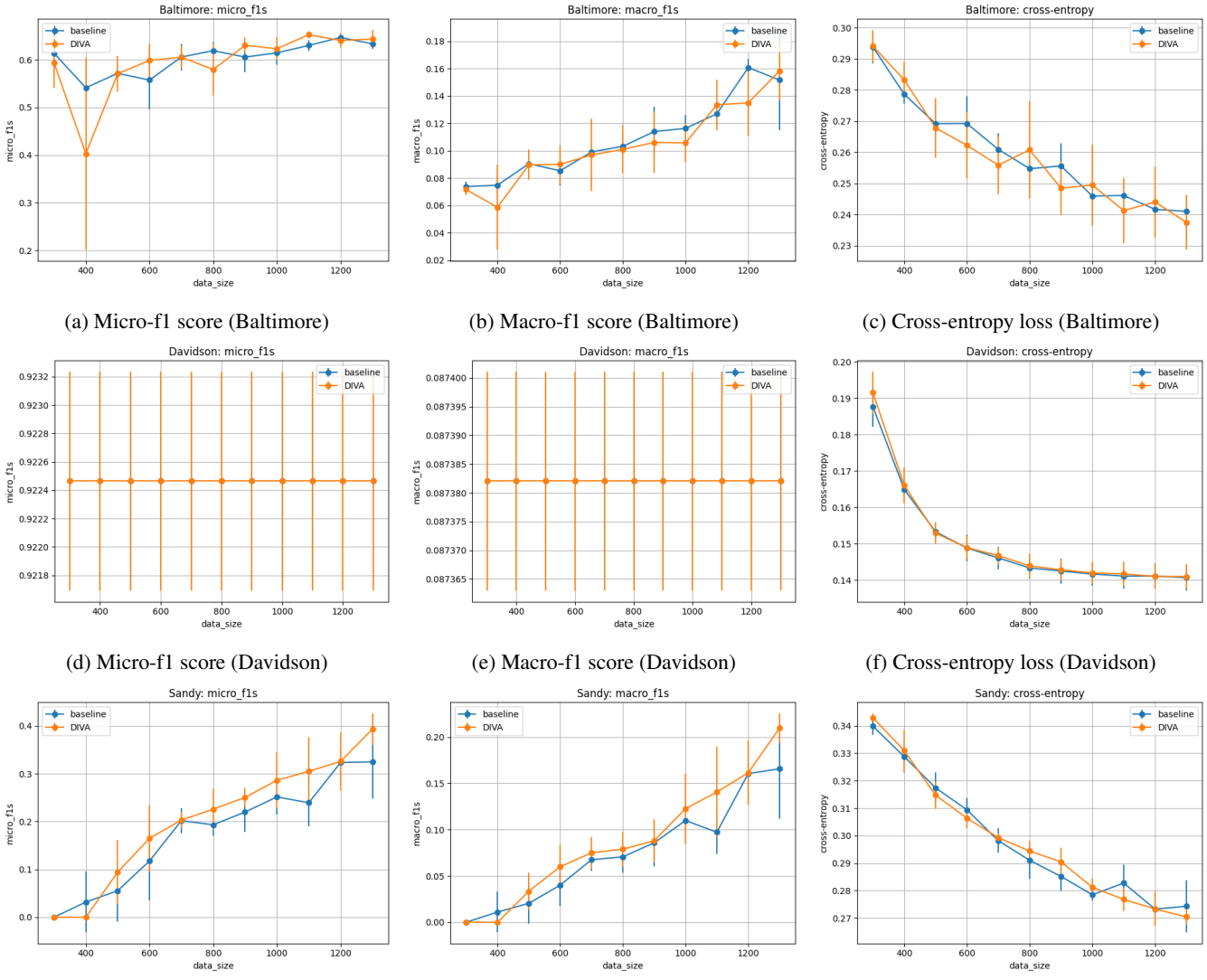


(c) Cross-entropy loss (BLM)

Figure 5: Results for different data selection strategy: Random, Uncertainty, Core-set selection

Table 5: Distribution of value labels over domains of MFTC

Value	ALM	Baltimore	BLM	Davidson	Election	MeToo	Sandy
Care	456	171	321	9	398	206	992
Harm	735	244	1037	138	588	433	793
Fairness	515	133	522	4	560	391	179
Cheating	505	519	876	62	620	685	459
Loyalty	244	373	523	41	207	322	415
Betrayal	40	621	169	41	128	366	146
Authority	244	17	276	20	169	415	443
Subversion	91	257	303	7	165	874	451
Purity	81	40	108	5	409	173	56
Degradation	122	28	186	67	138	941	91
Nonmoral	1744	3826	1583	4509	2501	1565	1313



(a) Micro-f1 score (Baltimore)

(b) Macro-f1 score (Baltimore)

(c) Cross-entropy loss (Baltimore)

(d) Micro-f1 score (Davidson)

(e) Macro-f1 score (Davidson)

(f) Cross-entropy loss (Davidson)

(g) Micro-f1 score (Sandy)

(h) Macro-f1 score (Sandy)

(i) Cross-entropy loss (Sandy)

Figure 6: Results for Baltimore, Davidson and Sandy domains