Combining Active Learning and Task Adaptation with BERT for Cost-Effective Annotation of Social Media Datasets

Jens Lemmens
University of Antwerp (CLiPS)
Prinsstraat 13
2000 Antwerp (Belgium)
jens.lemmens@uantwerpen.be

Walter Daelemans
University of Antwerp (CLiPS)
Prinsstraat 13
2000 Antwerp (Belgium)
walter.daelemans@uantwerpen.be

Abstract

Social media provide a rich source of data that can be mined and used for a wide variety of research purposes. However, annotating this data can be expensive, yet necessary for state-of-the-art pre-trained language models to achieve high prediction performance. Therefore, we combine pool-based active learning based on prediction uncertainty (an established method for reducing annotation costs) with unsupervised task adaptation through Masked Language Modeling (MLM). The results on three different datasets (two social media corpora, one benchmark dataset) show that task adaptation significantly improves results and that with only a fraction of the available training data, this approach reaches similar F1-scores as those achieved by an upper-bound baseline model fine-tuned on all training data. We hereby contribute to the scarce corpus of research on active learning with pre-trained language models and propose a cost-efficient annotation sampling and fine-tuning approach that can be applied to a wide variety of tasks and datasets.

1 Introduction

Approximately 59% of the population worldwide use social media (Chaffey, 2023). Collectively, they post more than half a million comments on Facebook each minute, and a grand total of 500 million tweets per day (Shepherd, 2023; Aslam, 2023). These statistics indicate that social media are a virtually inexhaustible source of data, and a large part of this data can be accessed for research purposes. However, annotating this data, which is often necessary to achieve high prediction performance with pre-trained language models, can be an expensive and time-consuming process. One approach that has been proposed in previous research to reduce annotation costs is active learning (AL), which aims at optimizing the annotation effort by selecting specific data points from an unlabeled data pool which are expected to contribute the most to a model's learning phase (Settles, 2009). Although AL has proven its usefulness throughout decades of research, it remains a data selection method, which makes it challenging to use as only tool for annotation cost reduction and to reach upper bound performance (achieved by supervised learning on all available data).

In this work, we therefore exploit the capabilities of transformer-based pre-trained language models to learn from unsupervised data through their pre-training task. Concretely, we combine AL based on the prediction uncertainty of a model with unsupervised task adaptation through masked language modeling (MLM) to investigate whether this combination of techniques allows reaching the upper bound and with how much data. We test this approach by using different AL protocols on three publicly available datasets (2 social media datasets and 1 benchmark) that are costly to annotate, e.g., because they contain many fine-grained labels or the task is difficult to learn with little data. The experiments presented in this work show that using task adaptation before AL has a significant effect on model performance, and that substantially less data is needed to reach upper bound performance, therefore reducing annotation expenses.

2 Related research

2.1 Active learning

Active learning, or "sequential sampling", has been studied since the 1990's (Lewis and Gale, 1994; Lewis and Catlett, 1994; Cohn et al., 1994, 1996). Originally, AL referred to prioritizing certain entries in an unlabeled dataset during the annotation process, based on the prediction confidence of a model trained on a small initial subset of labeled data, as shown in Figure 1. The intuition behind this sampling strategy is that the less certain a model is about a prediction, the more the model can learn from this data point during training, thereby

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 237–250
July 14, 2023 ©2023 Association for Computational Linguistics

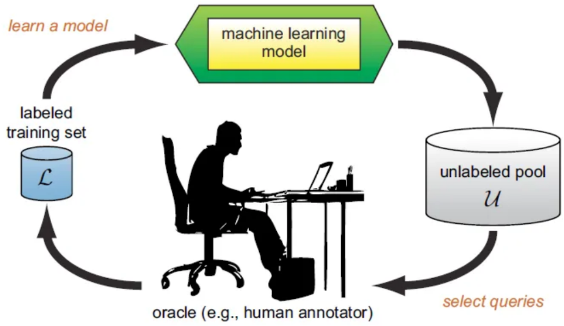

Figure 1: Illustration of a pool-based active learning process (Settles, 2009): A random sample is first labeled by an oracle (human annotator) and used to train a prototype model. This model then predicts the labels of the rest of the unlabeled data. Afterwards, the n data points with the lowest prediction confidence are annotated and used to update the model. This process continues until the annotation budget is depleted or until no more improvements are observed in the learning curve.

maximizing the return on annotation investment.

In the last decades, AL has shown improvements in different tasks and models, such as text classification with k-nearest neighbors (Shi et al., 2008), word sense disambiguation with support vector machines (Zhu and Hovy, 2007), and machine translation with recurrent neural networks (Vashistha et al., 2022). Although AL is most commonly based on prediction uncertainty, sampling can also be based on model disagreement, such as in BALD (Bayesian AL by Disagreement, (Houlsby et al., 2011)), gradient information, such as in BADGE (Batch AL by Diverse Gradient Embeddings, Ash et al. (2019)), typicality or density (Zhu et al., 2008), batch diversity or representativeness (Shi et al., 2021), and other metrics (or a combination of any of the aforementioned, Settles (2009)).

2.2 Active learning with language models

Although AL with pre-trained transformers has gained interest in recent years, the amount of research remains relatively scarce compared to AL with traditional machine learning or neural models. Existing work, e.g. Schröder et al. (2022), has examined the vanilla uncertainty-based query strategy for various binary and multi-class text classification experiments and shown that this strategy is also effective for pre-trained language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). Similarly, Ein-Dor et al. (2020) investigate the effect of annotation sample selection based on prediction uncertainty, expected gradient length and batch representativeness compared to random sample selection. They hereby focus on binary text classification tasks where the positive class is in the minority and show that all strategies perform substantially better than a random sampling strategy.

Recently, Rotman and Reichart (2022) were the first to explore multi-task AL with pre-trained language models, building upon the work of Reichart et al. (2008), who introduced the concept for traditional machine learning methods, and Ikhwantri et al. (2018), who used it for (non-pre-trained) neural architectures. Entropy-based confidence, both in isolation and when using dropout agreement, was used for multi-task AL, and compared to single-task AL and random sampling. Additionally, they investigated the effect of task weighting in ranking samples on informativeness. Their experiments showed that multi-task AL is an efficient way for annotation cost reduction, but that the precise method should depend on the task(s).

Finally, it is noteworthy that any model utilized in an AL setting can be trained in two ways: by updating it after each step, or by re-initializing it and training the entire model on all available annotated data (Schröder et al., 2022; Ein-Dor et al., 2020; Hu et al., 2018; Shen et al., 2017). In general, language models are more frequently re-initialized than updated, because they tend to be unstable when incrementally fine-tuned on low amounts of data, resulting in lower performance and higher standard deviations across different random seeds or hyperparameters (Dodge et al., 2020). However, the effect of re-initializing versus updating language models during AL is understudied. This work will therefore investigate whether re-initializing language models is indeed the preferred approach when using the standard uncertainty-based AL approach, and when combining it with task adaptation.

2.3 Task adaptation

As mentioned, this work combines AL with task adaptation. The latter refers to learning training data in an unsupervised manner before fine-tuning on it for a specific end task. For example, Buhmann et al. (2022) show that task adaptation has a positive effect on their question-answering model used for VaccinChat[1]: a user interface that answers Dutch-language user questions about the COVID-19 vac-

[1] https://vaccinchat.be/

cine, specifically for Flemish (Dutch speaking Belgian) users. Similarly, Mehri et al. (2020) show that task adaptation through performing MLM on the training data before fine-tuning increases the performance of their task-oriented dialogue system. In the experiments presented in this paper, we adopt the MLM approach as task adaptation step before commencing AL.

3 Methodology

3.1 Data

Existing AL research often relies on datasets that are inexpensive to annotate or tasks that are relatively easily learned by models, even when little data is available. Examples are the IMDB movie review, YELP polarity, SST-2, and TREC datasets for NLP (Maas et al., 2011; Zhang et al., 2015; Socher et al., 2013; Hovy et al., 2001), and the MNIST and Fashion-MNIST for computer vision (Deng, 2012; Xiao et al., 2017). This results in experiments where very small initial training samples are used, which are then increased in equally small steps, while still achieving relatively high prediction performances. Although it is necessary to create artificial AL setups, the aforementioned experimental settings are in our opinion inappropriate for research on AL, because the method is in reality the most effective when working with data that is expensive to annotate or when many examples are needed to gain high prediction accuracy.

Taking this into account, we use the FRENK (Ljubesic et al., 2019), and GoEmotions (Demszky et al., 2020) datasets for our experiments and validate the results on an additional benchmark: the 20 News Groups dataset (Lang, 1995). A detailed description of these datasets and why they are useful for AL experiments can be found below, and a summary of the statistics of each of these datasets can be found in Table 1. A fine-grained overview of the class distributions of the datasets can be found in Appendix A.

Table 1: Statistics of the data used in our experiments.

Dataset	Labels	Train	Val	Test
FRENK	4 (multi-class)	8,404	933	2,301
GoEmotions	28 (multi-label)	43,410	5,426	5,427
20 News Groups	20 (multi-class)	10,182	1,131	7,532

3.1.1 FRENK

The FRENK dataset[2] consists of Facebook comment threads on news item posts about two topics: migrants and the LGBTQ+ community (Ljubesic et al., 2019). Although the dataset contains Slovene and English comments, the current work only utilizes the English partition. The data contains labels concerning the topic (LGBTQ+ or migrants), the target of the hate speech (topic, related to topic, journalist/medium, other, no target), and the type of hate speech, which is the task we tackle in this study. FRENK distinguishes six types of hate speech in its annotation scheme:

1. **Background-violence** consists of messages that call for violence based on the personal background of the target (e.g., religion, gender, race or ethnicity).
2. **Other-violence** contains messages that call for violence for another reason than the background of the target, e.g., an opinion expressed by the target.
3. **Background-offensive** refers to messages that contain offensive statements that are aimed at the background of the target.
4. **Other-offensive** are messages that contain offensive language towards any aspect of the speaker but their personal background.
5. **Inappropriate speech** pertains to messages that contain vulgar and/or other types of offensive language that is not aimed towards a specific target (hence this category is technically not hate speech).
6. **Appropriate speech**, which does not contain any form of offensive or violent language.

Since the violent language classes contain very few entries, and the experiments in this paper require using small samples of training data, labels (1) and (2) were combined to form one "violent language" class, and labels (3) and (4) were used to form an "offensive language" class for the purpose of this paper, resulting in a total of 4 labels (the label distribution with these newly formed classes can be found in Table 15, Appendix A). The training, validation, and test partitions contain 8,404, 933 and 2,301 entries, respectively.

Since hate speech is a term that is open for interpretation and its identification depends on the personal and cultural background of the annotator,

[2] `https://huggingface.co/datasets/classla/FRENK-hate-en`

multiple annotators are needed to generate high quality labels and avoid bias (Sap et al., 2022). In addition, the labels in FRENK show strong class imbalances, which is why many comments are needed in order to collect sufficient annotations for the underrepresented classes, resulting in high annotation costs. Further, annotators were required to read the comments thoroughly, since the labels contain a hierarchy of importance in cases where multiple types of hate speech occur in one message (other < background; offensive language < violent language), which increases annotation time even more.

3.1.2 GoEmotions

The GoEmotions dataset[3] contains Reddit comments annotated with 28 emotions (incl. "neutral") in a multi-label setting (Demszky et al., 2020). The dataset is divided in 43,410 cases for training, 5,426 for validating, and 5,427 for testing.

Due to its high number of classes and multi-label scheme, annotating the GoEmotions dataset is labor-intensive. Given its large class imbalances, it is particularly difficult to gain a performance increase in the smaller classes by annotating more samples, since many samples need to be annotated before collecting a substantial amount of messages that express emotions that are infrequent in the dataset.

3.1.3 20 News Groups

The 20 News Groups dataset[4] contains approximately 20,000 news groups posts each associated with 1 out of 20 different topic classes (Lang, 1995). We use this benchmark with many fine-grained classes as an additional test for the proposed approach next to the above mentioned social media datasets, which are the focus of this paper.

3.2 Approaches

3.2.1 Baseline approaches

Random sampling For this baseline, the training data was sampled randomly so that the effect of the AL strategies could be measured.

Upper bound This approach refers to fine-tuning with all available training data in order to estimate the highest possible performance that can be achieved with standard fine-tuning.

3.2.2 AL approaches

Model re-initialization This method refers to the standard AL strategy for language models as proposed in (Schröder et al., 2022; Ein-Dor et al., 2020): An initial sample is used to fine-tune a model, which then predicts labels for the rest of the training data. A second sample is then selected based on the confidence of the model and a new model is initialized and fine-tuned using all annotated data. This process then repeats itself n times.

Checkpoint updating This method is identical to the approach above, with the only difference being that each time a new batch of annotated data is selected, the model is not re-initialized, but fine-tuning continues with the new annotated sample starting from the final checkpoint of the previous round of fine-tuning.

Two-step learning This approach is a specific form of AL where a model is first fine-tuned and then updated once. For the first fine-tuning stage, an initial random sample is used. After predicting the labels of the rest of the training data with this model, it is fine-tuned a second time using the top n most uncertain entries, where n is determined by the rest of the annotation budget. In contrast, "checkpoint updating" refers to annotating various batches of fixed size and updating the model after each batch (i.e., in more than two steps).

3.3 Experimental setup

The annotation process was replicated as follows: assuming that a labeled validation and test set are available, a random batch consisting of ca. 10% of all training data was used as an initial sample for fine-tuning. Then, the amount of training data was incrementally increased until approx. 50% of the available training data was used. For all classification experiments, a BERT-base-uncased model (Devlin et al., 2019) was fine-tuned for 5 epochs with a batch size of 32 and a learning rate of 5e-5. The model was evaluated after each epoch, and predictions on the test set were made with the checkpoint that yielded the highest macro-averaged F1-score on the validation set.

In experiments where task adaptation was used, the model was first fine-tuned on the entire unlabeled training dataset through MLM for the duration of 5 epochs with a batch size of 64 and learning rate of 1e-4. The model checkpoint with lowest validation loss was used for further experiments.

[3] https://huggingface.co/datasets/go_emotions

[4] https://huggingface.co/datasets/SetFit/20_newsgroups

Table 2: Macro-averaged results on **FRENK** using **random sampling**. The last row represents the upper bound baseline.

n	Pre	Rec	F1	Std
1,000	47.5	42.5	43.4	1.5
1,500	47.4	45.4	45.9	0.7
2,000	54.1	46.3	48.0	4.7
2,500	63.1	49.0	52.0	4.8
3,000	65.4	50.8	54.2	3.5
3,500	66.9	49.1	52.9	4.6
4,000	67.2	53.4	56.4	1.7
8,404	62.0	57.8	59.1	1.5

Table 3: Macro-averaged results on **FRENK** using AL with model **re-initialization**. Best F1 are in bold, experiments reaching the upper bound are in grey, and statistical significance is indicated with asterisks.

	Re-initialization				**+ Adaptation**			
n	**Pre**	**Rec**	**F1**	**Std**	**Pre**	**Rec**	**F1**	**Std**
1,000	47.5	42.5	43.4	1.5	50.2	44.4	**45.7****	2.1
1,500	50.2	45.8	46.9	4.6	57.5	46.9	**48.7**	3.8
2,000	63.7	45.8	48.5	1.8	59.8	49.2	**51.8****	3.3
2,500	63.4	50.9	54.5	5.0	65.6	56.0	**59.4***	1.5
3,000	62.8	51.8	55.0	3.1	65.3	56.5	**59.4**	2.1
3,500	63.9	54.1	57.1	2.6	66.2	56.0	**59.3**	2.8
4,000	65.0	53.7	57.4	2.3	63.3	59.6	**61.0****	3.2

Table 4: Macro-averaged results on **FRENK** using AL with model **updating**. Best F1 are in bold, experiments reaching the upper bound are in grey.

	Checkpoint updating				**+ Adaptation**			
n	**Pre**	**Rec**	**F1**	**Std**	**Pre**	**Rec**	**F1**	**Std**
1,000	47.5	42.5	43.4	1.5	50.2	44.4	**45.7**	2.1
1,500	49.6	45.6	**46.8**	4.0	55.7	45.3	46.5	0.9
2,000	51.6	47.7	48.3	3.5	56.9	48.3	**50.7**	5.3
2,500	54.4	49.8	50.7	3.5	65.8	49.0	**53.1**	1.9
3,000	56.7	47.9	49.7	4.3	64.9	51.6	**55.3**	1.1
3,500	57.2	52.4	53.8	4.7	60.8	52.9	**54.8**	5.4
4,000	62.7	53.3	**56.2**	2.8	61.8	52.4	55.2	2.9

Table 5: Macro-averaged results on **FRENK** using AL with **two-step learning**. Best F1 are in bold, experiments reaching the upper bound are in grey.

	Two-step				**+ Adaptation**			
n	**Pre**	**Rec**	**F1**	**Std**	**Pre**	**Rec**	**F1**	**Std**
1,000	47.5	42.5	43.4	1.5	50.2	44.0	**45.7**	2.1
1,500	49.6	45.6	**46.8**	4.0	55.7	45.3	46.5	0.9
2,000	54.7	46.5	48.5	4.3	69.5	51.4	**55.6**	2.7
2,500	55.4	48.0	50.2	5.3	64.1	53.6	**56.7**	2.2
3,000	62.9	48.9	52.4	5.0	65.3	51.7	**55.6**	2.1
3,500	61.9	54.4	55.8	3.0	63.8	56.7	**58.5**	1.7
4,000	64.2	52.1	55.5	4.5	67.3	55.4	**59.3**	3.3

4 Results

4.1 FRENK

Table 2 shows the random and upper bound baselines on the FRENK dataset, whereas Table 3, 4, and 5 show the results achieved for AL with model re-initialization, checkpoint updating, and two-step learning, respectively. The upper bound F1-macro amounts to 59.1%, which was surpassed when using 2,500 training instance in an AL fine-tuning process with model re-initialization after task adaptation. Specifically, using 29.7% of the available labeled training data led to a performance of 59.4% F1-macro. When increasing the training data to 4,000 entries, 47.6% of all available training data, the F1-score reached 61.0%.

Overall, model re-initialization when fine-tuning after querying a new sample of annotated data achieves best results, followed by two-step learning, and finally incremental checkpoint updating. In the case of the latter, we even observe a small decrease in performance (from 55.3% to 54.8%) and high standard deviation of the F1-score (5.4%) when updating the model with 500 unseen entries after it was already fine-tuned cumulatively on 3,000 entries. This shows that updating model checkpoints with smaller samples of data is inefficient, cf. Schröder et al. (2022); Ein-Dor et al. (2020). For the best AL approach, i.e. model re-initialization, a McNemar test was conducted for each sample size to determine whether the improvements after task adaptation are statistically significant (McNemar, 1947), as shown in Table 3. These tests indicate that the improvements in 4 out of 7 experiments are statistically significant, while the other three still show substantial improvements.

To illustrate where most improvements were made after using task adaptation, Table 19 (Appendix C), shows the results per class after fine-tuning on 2,500 training entries using different sampling approaches (random sampling, standard AL, AL after task adaptation). In this table, it can be observed that the highest improvements were made in the most difficult and underrepresented class, namely the "violent speech" class. More precisely, results improved from 18.3% to 32.3% when using standard AL instead of random sampling, and to 47.1% after using task adaptation before AL. Noteworthy is that the performance for the "inappropriate speech" class dropped with 1.4% when using standard AL (compared to random sampling), but when using task adaptation, this performance drop was no longer observed.

Table 6: Macro-averaged results on **GoEmotions** using **random sampling**. The last row represents the upper bound baseline.

n	Pre	Rec	F1	Std
4,000	12.8	8.7	9.7	2.6
6,000	18.2	13.0	14.2	1.4
8,000	31.3	18.7	21.5	0.9
10,000	37.7	22.8	26.3	1.5
12,000	48.9	28.1	32.7	2.5
14,000	51.1	32.5	37.5	1.4
16,000	54.6	34.7	39.7	0.4
18,000	52.2	36.1	40.9	0.4
20,000	52.1	37.1	41.8	0.9
22,000	52.4	38.7	43.1	0.8
43,410	54.4	43.7	47.2	0.7

Table 7: Macro-averaged results on **GoEmotions** using AL with model **re-initialization**. Best F1 are in bold, experiments reaching the upper bound are in grey. See Appendix B, Table 18 for statistical significance per class / sample size.

	Re-initialization				+ Adaptation			
n	**Pre**	**Rec**	**F1**	**Std**	**Pre**	**Rec**	**F1**	**Std**
4,000	12.8	8.7	9.7	2.6	15.5	11.9	**12.9**	0.4
6,000	22.0	15.3	17.0	0.9	36.3	21.8	**24.7**	0.5
8,000	36.6	22.6	25.8	1.4	47.6	30.0	**34.3**	0.5
10,000	49.0	29.6	34.2	0.7	51.2	34.2	**39.3**	0.6
12,000	52.0	33.0	38.1	0.9	54.5	36.8	**41.9**	0.8
14,000	54.7	35.5	40.8	0.6	56.9	39.2	**44.2**	0.4
16,000	56.6	38.1	43.5	0.6	55.2	40.5	**45.2**	0.6
18,000	55.5	38.9	43.9	0.7	55.2	40.9	**45.5**	0.1
20,000	55.7	40.3	45.1	0.5	54.6	41.8	**46.2**	0.5
22,000	55.7	41.0	45.6	0.8	54.4	42.5	**46.3**	0.4

Table 8: Macro-averaged results on **GoEmotions** using AL with **checkpoint updating**. Best F1 are in bold, experiments reaching the upper bound are in grey.

	Checkpoint updating				+ Adaptation			
n	**Pre**	**Rec**	**F1**	**Std**	**Pre**	**Rec**	**F1**	**Std**
4,000	12.8	8.7	9.7	2.6	14.1	11.6	**12.6**	0.2
6,000	22.4	16.2	17.9	1.7	29.7	20.7	**22.9**	2.1
8,000	33.8	21.6	24.7	1.7	42.4	25.7	**29.1**	3.2
10,000	44.1	26.6	30.5	1.9	50.9	30.3	**34.9**	2.0
12,000	51.7	29.9	34.9	1.8	53.5	35.7	**40.6**	1.5
14,000	56.5	33.8	39.4	1.0	55.9	36.8	**42.0**	0.7
16,000	57.2	34.7	40.6	1.2	55.9	37.7	**42.7**	0.7
18,000	56.2	34.4	41.1	0.7	54.5	39.1	**43.6**	0.4
20,000	56.1	37.6	43.1	0.9	55.0	38.8	**43.6**	0.7
22,000	55.9	37.7	43.0	0.8	54.8	39.3	**43.7**	0.6

Table 9: Macro-averaged results on **GoEmotions** using AL with **two-step learning**. Best F1 are in bold, experiments reaching the upper bound are in grey.

	Two-step				+ Adaptation			
n	**Pre**	**Rec**	**F1**	**Std**	**Pre**	**Rec**	**F1**	**Std**
4,000	12.8	8.7	9.7	2.6	14.1	11.6	**12.6**	2.5
6,000	22.4	16.2	17.9	1.7	29.7	20.7	**22.9**	2.1
8,000	25.0	19.2	20.9	2.6	38.6	24.6	**27.5**	2.4
10,000	35.1	22.6	24.8	2.5	41.4	28.3	**31.4**	2.5
12,000	45.8	26.1	29.3	0.9	46.9	30.5	**34.2**	2.2
14,000	46.9	28.9	32.7	0.8	47.8	33.3	**37.3**	2.0
16,000	50.0	32.0	36.1	1.6	49.9	35.4	**39.6**	2.8
18,000	52.2	35.3	39.6	1.8	51.8	37.3	**41.3**	1.4
20,000	53.3	36.0	40.7	1.2	54.0	39.1	**42.8**	1.6
22,000	53.2	39.0	43.3	1.1	55.4	40.6	**44.6**	0.2

4.2 GoEmotions

The results of the GoEmotions dataset are shown in Table 6 (random and upper bound baselines), 7 (model re-initialization), 8 (model updating), and 9 (two-step learning). As shown, the upper bound baseline achieves an F1-macro score of 47.2% on the test set (using all 43,410 samples for fine-tuning). In contrast with the experiments on the FRENK dataset, we observe that the upper bound baseline could not be matched with the utilized sample sizes. Nevertheless, it can be observed that task adaptation improves results for all AL approaches and sample sizes. Similarly to the results of the FRENK dataset, model re-initialization yields better results than checkpoint updating (regardless of whether this is done incrementally or through two-step learning). In the case of AL by checkpoint updating, we even observe a decrease in performance compared to random sampling when using 22,000 entries cumulatively for fine-tuning. This evidences the inefficiency of incremental fine-tuning of pre-trained language models on small data samples. Statistical significance for the experiments with model re-initialization in this multi-label setting was determined by conducting a McNemar test for each individual class for each sample size. A summary of these tests can be found in Appendix B, Table 18. The three emotions where statistically significant improvements were observed the most frequently were "approval", "confusion", and "amusement", and in total, 6 out of 28 emotions were never predicted more accurately with statistical significance.

Finally, Table 20 (Appendix C) shows that the improvements are primarily found in the most difficult classes, namely those that were not predicted in the experiments with random sampling: "annoyance", "caring", "confusion", "desire", "disappointment", "excitement", "fear" and "surprise". The bulk of these did also not get predicted when using standard AL, but improvements could be observed when using task adaptation.

Table 10: Macro-averaged results on **20 News Groups** using **random sampling**. The last row represents the upper bound baseline.

n	Pre	Rec	F1	Std
1,000	56.7	55.4	51.8	0.9
1,500	59.6	60.0	58.0	1.0
2,000	63.1	63.4	62.4	0.5
2,500	64.0	63.8	63.1	1.0
3,000	65.4	64.7	64.3	0.7
3,500	66.0	65.6	65.3	0.3
4,000	66.1	65.5	65.2	0.3
4,500	66.8	66.1	65.9	0.4
5,000	67.4	66.5	66.5	0.4
5,500	68.0	67.1	67.1	0.4
10,182	69.8	68.9	69.1	0.5

Table 11: Macro-averaged results on **20 News Groups** using AL with model **re-initialization**. Best F1 are in bold, experiments reaching the upper bound are in grey.

	Re-initialization				+ Adaptation			
n	Pre	Rec	F1	Std	Pre	Rec	F1	Std
1,000	56.7	55.4	51.8	0.9	62.2	60.9	**58.3*** **	0.7
1,500	56.9	57.6	54.7	1.3	62.8	62.2	**60.2*** **	0.8
2,000	59.8	60.6	58.9	1.3	64.6	64.2	**63.4*** **	1.1
2,500	62.7	62.2	61.4	1.5	66.5	65.6	**65.3*** **	0.5
3,000	64.4	63.9	63.5	0.8	67.4	66.6	**66.5*** **	0.5
3,500	65.9	65.2	65.1	0.5	68.3	67.5	**67.4*** **	0.2
4,000	66.8	66.2	66.1	0.5	68.5	67.8	**67.7*** **	0.3
4,500	67.4	66.4	66.3	0.6	68.8	68.1	**68.0*** **	0.9
5,000	68.1	67.3	67.3	0.4	69.4	68.6	**68.7*** **	0.9
5,500	68.8	68.0	68.1	0.4	69.9	69.0	**69.2*** **	0.2

Table 12: Macro-averaged results on **20 News Groups** using AL with model **updating**. Best F1 are in bold, experiments reaching the upper bound are in grey.

	Checkpoint updating				+ Adaptation			
n	Pre	Rec	F1	Std	Pre	Rec	F1	Std
1,000	56.7	55.4	51.8	0.9	62.2	60.9	**58.3**	0.7
1,500	54.3	52.7	50.0	3.7	62.1	58.5	**56.4**	1.0
2,000	57.1	54.9	54.4	2.3	62.7	61.3	**60.0**	2.3
2,500	59.5	58.5	58.0	0.8	64.1	63.3	**62.4**	1.2
3,000	63.8	60.2	60.1	1.5	66.0	64.1	**63.4**	1.5
3,500	63.7	62.2	63.3	1.3	67.0	65.5	**65.7**	0.9
4,000	64.9	63.2	64.2	0.5	67.7	66.4	**66.4**	0.5
4,500	66.4	64.4	65.2	0.5	67.6	66.5	**66.5**	1.0
5,000	66.5	65.2	65.5	0.2	68.3	66.6	**66.7**	0.8
5,500	66.7	64.7	**66.3**	0.3	67.4	65.5	65.2	1.7

Table 13: Macro-averaged results on **20 News Groups** using AL with **two-step learning**. Best F1 are in bold, experiments reaching the upper bound are in grey.

	Two-step				+ Adaptation			
n	Pre	Rec	F1	Std	Pre	Rec	F1	Std
1,000	56.7	55.4	51.8	0.9	62.2	60.9	**58.3**	0.7
1,500	54.3	52.7	50.0	3.7	62.1	58.5	**56.4**	1.0
2,000	59.0	56.4	54.4	2.3	63.8	61.6	**60.6**	1.5
2,500	59.9	59.3	58.0	0.8	64.9	62.8	**62.3**	1.6
3,000	62.5	60.9	60.1	1.5	65.8	64.6	**64.4**	1.6
3,500	64.6	63.4	63.3	1.3	67.2	65.9	**65.0**	0.9
4,000	65.2	64.3	64.2	0.5	67.5	66.4	**66.3**	0.9
4,500	66.2	65.3	65.2	0.5	68.1	66.9	**67.1**	0.5
5,000	66.5	65.6	65.5	0.2	68.6	67.5	**67.6**	0.3
5,500	67.1	66.3	66.3	0.3	69.0	68.0	**68.2**	0.4

4.3 20 News Groups

As shown in Table 10, the upper bound F1-macro score achieved on the 20 News Groups dataset is 69.1%. This score was (only) achieved when using the AL protocol with model re-initialization and task adaptation after sampling 5,500 training entries, which equals 54.0% of the available training data.

Similarly to the experiments on the other datasets, task adaptation improves all three explored AL protocols, although model re-initialization remains the best of the three. Interestingly, however, AL without task adaptation yields lower results than random sampling. For AL with model re-initialization, for example, there is an average decrease of 0.8% in F1-macro across all sample sizes. This shows that the standard uncertainty-based AL makes worse sampling choices than random sampling, and that uncertainty is therefore a suboptimal metric for measuring informativeness in this particular dataset. A possible explanation for this observation is that language models are often ill-calibrated and tend to be overconfident, even if their prediction is false (Yuan et al., 2020; Park and Caragea, 2022). Additionally, low prediction confidence may indicate that an entry is noisy, not just difficult to predict. A qualitative analysis of the entries that are selected early in the AL process shows that this is the case for the 20 News Groups dataset: many sampled posts contain merely a few words that are irrelevant to the topic, whereas other posts are lengthy and discuss a multitude of (irrelevant) topics causing the low prediction confidence in the classifier. Since data selection based on prediction uncertainty collects more noise than a random selection of data in this case, it prevents the model from learning useful information, especially in earlier samples.

After task adaptation, however, there is an average improvement of 2.6% over the random baseline, and an average gain of 3.2% over standard AL (improvements for all sample sizes are statistically significant). This shows that although sample selection based on model prediction uncertainty is suboptimal on some datasets, task adaptation can act as a safety net to avoid performance impairment due to the use of suboptimal metrics for measuring informativeness.

Finally, and similarly to the results observed on the other datasets, the highest performance increases are observed in the most difficult classes. For example, "talk.religion.misc" improves from 8.9% (random sampling) to 22.4% (AL with model re-initialization and task adaptation), as shown in Appendix C (Table 21).

The results of the experiments until this point have shown that task adaptation has a positive effect on AL: Significant improvements in F1-score could be observed, and in two datasets, the upper bound could be reached with a fraction of the annotations, while still showing substantial gains in the third dataset. In the case of 20 News Groups, we observed that traditional AL had a negative effect on model performance, although task adaptation countered this effect. Finally, the results indicate that AL with model re-initialization leads to more stable fine-tuning than with model updating.

4.4 Ablation study

In this section, we investigate the effect of task adaptation in isolation, and whether using AL still has beneficial effects on model performance after task adaptation. In order to gain insights into this matter, task adaptation was used without AL, i.e. with random sample selection. The result of this experiment was then compared to the result achieved with random sampling, standard AL, and task adaptation combined with AL (as reported in the previous subsections).

The results of the above mentioned experiments can be found in Table 14: For FRENK, it can be concluded that task adaptation alone does not improve results when using random sample selection, although using task adaptation and AL leads to improvements of 4.6% on average. In comparison, standard AL leads to average improvements of 1.7%. This surprising result indicates that BERT does not learn new knowledge from task adaptation, but that this technique causes better sample selection during the AL stage. There may be different reasons why task adaptation has less effect on FRENK than on the other datasets. For example, the data in FRENK could resemble the data used to pre-train BERT more than is the case for the other datasets, so that less new information is

Table 14: Improvements (F1) of task adaptation and AL over random sampling across all sample sizes (first sample size was not included in the calculations for the standard AL experiments, since this experiment is identical to that of random sample selection).

FRENK					GoEmotions					20 News Groups				
n	**Random**	**Random + adaptation**	**AL**	**AL + adaptation**	**n**	**Random**	**Random + adaptation**	**AL**	**AL + adaptation**	**n**	**Random**	**Random + adaptation**	**AL**	**AL + adaptation**
1,000	43.4	**45.7 (+2.3)**	43.4 (+0.0)	**45.7 (+2.3)**	4,000	9.7	**12.9 (+3.2)**	9.7 (+0.0)	**12.9 (+3.2)**	1,000	51.8	**58.3 (+7.5)**	51.8 (+0.0)	**58.3 (+7.5)**
1,500	45.9	46.8 (+0.9)	46.9 (+1.0)	**48.7 (+2.8)**	6,000	14.2	19.9 (+5.7)	17.0 (+2.8)	**24.7 (+10.5)**	1,500	58.0	**62.7 (+4.7)**	54.7 (-3.3)	60.2 (+2.2)
2,000	48.0	48.3 (+0.3)	48.5 (+0.5)	**51.8 (+3.8)**	8,000	21.5	28.6 (+7.1)	25.8 (+4.3)	**34.3 (+12.8)**	2,000	62.4	**64.9 (+2.5)**	58.9 (-4.0)	63.4 (+1.0)
2,500	52.0	50.0 (-2.0)	54.5 (+2.5)	**59.4 (+7.4)**	10,000	26.3	35.9 (+9.6)	34.2 (+7.9)	**39.3 (+13.0)**	2,500	63.1	**66.3 (+3.2)**	61.4 (-1.7)	65.3 (+2.3)
3,000	54.2	53.2 (-1.0)	55.0 (+0.8)	**59.4 (+5.2)**	12,000	32.7	38.9 (+6.2)	38.1 (+5.4)	**41.9 (+9.2)**	3,000	64.3	**67.2 (+2.9)**	63.5 (-0.8)	66.5 (+2.2)
3,500	52.9	52.6 (-0.3)	57.1 (+4.2)	**59.3 (+6.4)**	14,000	37.5	40.5 (+3.0)	40.8 (+3.3)	**44.2 (+7.3)**	3,500	65.3	**67.5 (+2.2)**	65.1 (-0.2)	67.4 (+2.1)
4,000	56.4	56.2 (-0.2)	57.4 (+1.0)	**61.0 (+4.6)**	16,000	39.7	41.2 (+1.5)	43.5 (+3.8)	**45.2 (+5.5)**	4,000	65.2	67.4 (+2.2)	66.1 (+0.9)	**67.7 (+2.5)**
Avg. improvement		+0.0	+1.7	**+4.6**	18,000	40.9	43.3 (+2.4)	43.9 (+3.0)	**45.5 (+3.0)**	4,500	65.9	67.9 (+2.0)	66.3 (+0.4)	**68.0 (+2.1)**
					20,000	41.8	43.9 (+2.1)	45.1 (+4.0)	**46.2 (+4.4)**	5,000	66.5	68.0 (+1.5)	67.3 (+0.8)	**68.7 (+2.2)**
					22,000	43.1	44.5 (+1.4)	45.6 (+2.5)	**46.3 (+3.2)**	5,500	67.1	67.5 (+0.4)	68.1 (+1.0)	**69.2 (+2.1)**
					Avg. improvement		+4.2	+4.1	**+7.2**	**Avg. improvement**		**+2.9**	-0.8	+2.6

being learned. Determining the exact cause of this observation, however, is beyond the scope of this paper and can be the subject of future work.

With respect to the GoEmotions dataset, which yielded the lowest F1-scores compared to FRENK and 20 News Groups, the improvements of using both task adaptation and AL was the highest of all datasets: 7.2% on average. Individually, task adaptation and AL showed improvements similar to each other, namely 4.2% and 4.1%, respectively.

Regarding 20 News Groups, we observe that standard active learning has a negative effect on performance overall, especially in the smaller sample sizes, as mentioned before. Task adaptation, however, shows improvements of 2.9% and 2.6%, without and with the additional use of AL, respectively. This shows that when active learning has a negative effect on performance, task adaptation can negate this effect and still improve upon standard fine-tuning with random data selection.

5 Conclusion

Main findings The current work combines uncertainty-based AL with task adaptation in order to learn from the data that could not be labeled due to limited annotation budget. It investigates the following research questions:

- What is the the effect of task adaptation on AL?
- How much data is needed to achieve the same performance with the proposed approach as with standard fine-tuning on all data?

The results of the experiments conducted on multiple datasets provide the following answers to these questions:

- Task adaptation has a significantly positive effect on AL, regardless whether the model is re-initialized or updated (although re-initialization consistently leads to better results). An analysis of the performance per class shows that the improvements are highest in the most difficult or underrepresented classes, and that the most difficult tasks in general show the highest improvements.
- In our experiments, 29.7% and 54% of all annotated data in the FRENK and 20 News Groups dataset, respectively, was needed to achieve the upper bound baseline. Although the proposed approach did not reach the upper bound baseline in the GoEmotions dataset with 50% or less of the training data, these results show that the proposed combination of approaches can lower annotation costs substantially.

Future research directions As mentioned earlier, AL with pre-trained language models remains understudied, in spite of it being an efficient annotation cost reduction method. Future research directions for AL with language models may include investigating the effect of calibration quality on uncertainty-based AL. An additional direction worth investigating is combining AL with other methods, such as data augmentation, weak supervision and domain adaptation, which are until now topics that are studied more extensively in other machine learning fields, such as computer vision (Zhao et al., 2020; Biegel et al., 2021; Xie et al., 2021; Hao et al., 2021; Zhan et al., 2022).

Limitations

The current work focuses on AL with pre-trained language models based on lowest prediction confidence. In spite of the effectiveness of the strategy shown both in these experiments and in previous work (Schröder et al., 2022; Ein-Dor et al., 2020), neural models are often not calibrated well (Yuan et al., 2020; Park and Caragea, 2022), which implies that the output of the softmax function could be a suboptimal metric for measuring prediction confidence, i.e. informativeness, for a given training sample. Future work on this topic should therefore investigate whether other metrics work better for AL with pre-trained language models and whether these metrics also benefit from unsupervised task adaptation. Additionally, experiments could only be conducted on a limited amount of tasks and datasets. Future work should shed new light on the usefulness of the proposed approach in different settings.

Acknowledgements

This research was funded by the Flemish government under FWO IRI project CLARIAH-VL and under the Research Program Artificial Intelligence Flanders.

References

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

Salman Aslam. 2023. Twitter by the numbers: Stats, demographics fun facts. [Accessed 17 April 2023].

Samantha Biegel, Rafah El-Khatib, Luiz Otávio Vilas Boas Oliveira, Max Baak, and Nanne Aben. 2021. Active WeaSuL: Improving weak supervision with active learning. *CoRR*, abs/2104.14847.

Jeska Buhmann, Maxime De Bruyn, Ehsan Lotfi, and Walter Daelemans. 2022. Domain- and task-adaptation for VaccinChatNL, a Dutch COVID-19 FAQ answering corpus and classification model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3539–3549, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Dave Chaffey. 2023. Global social media statistics research summary 2023. SmartInsights. [Accessed 17 April 2023].

David Cohn, Les Atlas, and Richard Ladner. 1994. Improving generalization with active learning. *Machine Learning*, 15(2):201–221.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active learning with statistical models. *CoRR*, cs.AI/9603104.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Li Deng. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

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

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *CoRR*, abs/2002.06305.

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

Ruqian Hao, Khashayar Namdar, Lin Liu, and Farzad Khalvati. 2021. A transfer learning–based active learning framework for brain tumor classification. *Frontiers in Artificial Intelligence*, 4.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2001. Toward semantics-based answer pinpointing. In *Proceedings of the First International Conference on Human Language Technology Research*.

Peiyun Hu, Zachary C. Lipton, Anima Anandkumar, and Deva Ramanan. 2018. Active learning with partial feedback. *CoRR*, abs/1802.07427.

Fariz Ikhwantri, Samuel Louvan, Kemal Kurniawan, Bagas Abisena, Valdi Rachman, Alfan Farizki Wicaksono, and Rahmad Mahendra. 2018. Multi-task active learning for neural semantic role labeling on low resource conversational corpus. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 43–50, Melbourne. Association for Computational Linguistics.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In Armand Prieditis and Stuart Russell, editors, *Machine Learning Proceedings 1995*, pages 331–339. Morgan Kaufmann, San Francisco (CA).

David D. Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In William W. Cohen and Haym Hirsh, editors, *Machine Learning Proceedings 1994*, pages 148–156. Morgan Kaufmann, San Francisco (CA).

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. *CoRR*, abs/cmp-lg/9407020.

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. *CoRR*, abs/1907.11692.

Nikola Ljubesic, Darja Fiser, and Tomaz Erjavec. 2019. The FRENK datasets of socially unacceptable discourse in slovene and english. *CoRR*, abs/1906.02045.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tür. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *CoRR*, abs/2009.13570.

Seo Yeon Park and Cornelia Caragea. 2022. On the calibration of pre-trained language models using mixup guided by area under the margin and saliency. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5364–5374, Dublin, Ireland. Association for Computational Linguistics.

Roi Reichart, Katrin Tomanek, Udo Hahn, and Ari Rappoport. 2008. Multi-task active learning for linguistic annotations. In *Proceedings of ACL-08: HLT*, pages 861–869, Columbus, Ohio. Association for Computational Linguistics.

Guy Rotman and Roi Reichart. 2022. Multi-task active learning for pre-trained transformer-based models.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 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*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2194–2203, Dublin, Ireland. Association for Computational Linguistics.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

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

Jack Shepherd. 2023. 30 essential facebook statistics you need to know in 2023. TheSocialShepherd. [Accessed 17 April 2023].

Shuo Shi, Yuhai Liu, Yuehua Huang, Shihua Zhu, and Yong Liu. 2008. Active learning for kNN based on bagging features. In *2008 Fourth International Conference on Natural Computation*, volume 7, pages 61–64.

Tianze Shi, Adrian Benton, Igor Malioutov, and Ozan İrsoy. 2021. Diversity-aware batch active learning for dependency parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2616–2626, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Neeraj Vashistha, Kriti Singh, and Ramakant Shakya. 2022. Active learning for neural machine translation.

Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747.

Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, Xinjing Cheng, and Guoren Wang. 2021. Active learning for domain adaptation: An energy-based approach. *CoRR*, abs/2112.01406.

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

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

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. *arXiv:1509.01626 [cs]*.

Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. 2020. Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1796–1806, Online. Association for Computational Linguistics.

Jingbo Zhu and Eduard Hovy. 2007. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 783–790, Prague, Czech Republic. Association for Computational Linguistics.

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

A Label distribution per dataset

Table 15: Class distribution of the FRENK dataset after combining the violent and offensive classes.

Class name	Train	Val	Test
Violent language	96	11	15
Offensive language	1,487	165	477
Inappropriate language	1,490	165	410
Acceptable language	5,331	592	1,399

Table 16: Class distribution of the 20 News Groups dataset.

Class name	Train	Val	Test
alt.atheism	534	59	394
comp.graphics	538	60	398
comp.os.ms-windows.misc	508	56	376
comp.sys.ibm.pc.hardware	535	60	396
comp.sys.mac.hardware	532	59	393
comp.windows.x	520	58	385
misc.forsale	526	59	390
rec.autos	418	47	310
rec.motorcycles	526	58	389
rec.sport.baseball	491	55	364
rec.sport.hockey	534	59	395
sci.crypt	535	59	396
sci.electronics	539	60	398
sci.med	339	38	251
sci.space	535	59	396
soc.religion.christian	540	60	399
talk.politics.guns	531	59	392
talk.politics.mideast	432	48	319
talk.politics.misc	532	59	394
talk.religion.misc	537	60	397

Table 17: Class distribution of the GoEmotions dataset.

Class name	id	Train	Val	Test
Admiration	0	4,130	488	504
Amusement	1	2,328	303	252
Anger	2	1,567	195	197
Annoyance	3	2,470	303	286
Approval	4	2,939	397	318
Caring	5	1,087	153	114
Confusion	6	1,368	152	139
Curiosity	7	2,191	248	233
Desire	8	641	77	74
Disappointment	9	1,269	163	127
Disapproval	10	2,022	292	220
Disgust	11	793	97	84
Embarrassment	12	303	35	30
Excitement	13	853	96	84
Fear	14	596	90	74
Gratitude	15	2,662	358	288
Grief	16	77	13	6
Joy	17	1,452	172	116
Love	18	2,086	252	169
Nervousness	19	164	21	16
Optimism	20	1,581	209	120
Pride	21	111	15	8
Realization	22	1110	127	109
Relief	23	153	18	7
Remorse	24	545	68	46
Sadness	25	1,326	143	108
Surprise	26	1,060	129	92
Neutral	27	14,219	1,766	1,606

B GoEmotions significance per class

Table 18: Indices of emotions that were predicted significantly more accurately when using task adaptation before AL with model re-initialization (per sample size).

n	*	**	***
4,000	-	-	0, 1, 18
6,000	20	26	1, 2, 4, 7, 14, 17, 25
8,000	1, 8, 9, 26	24	13
10,000	6, 10, 24, 25, 26	27	4
12,000	2, 13, 20	-	4
14,000	14, 27	3, 6, 10	4, 12
16,000	6	9, 12, 22	4
18,000	3, 5, 6, 12, 14	-	4, 22
20,000	2, 6	4, 22	-
22,000	1, 22	-	-

C Results per class

Table 19: Comparison of results per class on the FRENK dataset after fine-tuning on 4,000 entries using different sampling approaches. The reported results are averaged across 5 runs with different random seeds. The best results are in bold.

	Random			**Re-initialization**			**+ Adaptation**		
Class name	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Acceptable speech	76.0	87.7	81.4	76.2	88.5	81.8	77.4	87.4	**82.1**
Offensive speech	56.4	44.5	49.5	58.1	46.7	51.3	59.5	50.3	**54.1**
Violent speech	73.3	10.7	18.3	54.0	24.0	32.3	61.2	38.7	**47.1**
Inappropriate speech	63.2	47.3	54.0	65.3	44.2	52.6	64.3	47.3	**54.3**

Table 20: Comparison of results per class on the GoEmotions dataset after fine-tuning on 22,000 entries using different sampling approaches. The reported results are averaged across 5 runs with different random seeds. The best results are in bold.

	Random			**Re-initialization**			**+ Adaptation**		
Class	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Admiration	71.8	60.8	65.8	70.2	63.8	66.8	68.3	66.0	**67.0**
Amusement	80.9	81.4	81.2	81.5	80.1	80.7	80.4	84.9	**82.6**
Anger	67.9	10.3	17.4	64.5	27.5	37.8	61.3	33.4	**43.0**
Annoyance	0.0	0.0	0.0	0.0	0.0	0.0	24.3	1.2	**2.4**
Approval	58.1	23.0	32.9	32.6	2.6	4.8	51.4	25.3	**33.2**
Caring	0.0	0.0	0.0	0.0	0.0	0.0	59.9	17.8	**26.0**
Confusion	0.0	0.0	0.0	0.0	0.0	0.0	47.2	4.8	**8.7**
Curiosity	58.2	36.7	**44.9**	45.2	24.8	31.5	54.8	35.8	43.3
Desire	0.0	0.0	0.0	63.7	9.4	15.9	67.1	20.2	**30.3**
Disappointment	0.0	0.0	0.0	0.0	0.0	0.0	53.7	4.9	**8.5**
Disapproval	38.5	6.7	10.8	34.5	3.1	5.4	46.9	19.4	**27.4**
Disgust	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Embarrassment	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Excitement	0.0	0.0	0.0	20.0	0.6	1.1	79.5	19.2	**30.5**
Fear	0.0	0.0	0.0	66.5	33.8	43.9	67.5	63.3	**65.3**
Gratitude	93.4	88.0	90.6	92.8	89.1	90.9	93.0	89.3	**91.1**
Grief	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Joy	73.2	30.8	42.8	72.4	43.5	53.9	69.1	48.7	**57.1**
Love	82.0	77.0	79.4	81.1	80.5	**80.8**	82.0	78.7	80.3
Nervousness	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Optimism	70.7	31.9	43.8	69.5	38.8	49.5	67.0	42.2	**51.7**
Pride	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Realization	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Relief	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Remorse	35.7	4.3	7.0	49.3	30.0	34.4	63.0	52.1	**56.9**
Sadness	77.0	14.6	23.6	74.4	37.3	49.2	68.1	40.4	**50.6**
Surprise	0.0	0.0	0.0	41.6	8.7	13.0	59.4	33.9	**42.8**
Neutral	67.7	56.6	61.6	65.9	59.2	**62.4**	67.4	57.3	61.8

Table 21: Comparison of results per class on the 20 News Groups dataset after fine-tuning on 4,500 entries using different sampling approaches. The reported results are averaged across 5 runs with different random seeds. The best results are in bold

	Random			**Re-initialization**			**+ Adaptation**		
Class name	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
alt.atheism	39.3	57.0	**46.4**	44.0	47.0	45.1	46.5	47.0	46.1
comp.graphics	64.0	68.2	65.9	66.4	68.7	67.5	66.6	70.7	**68.4**
comp.os.ms-windows.misc	59.7	63.2	61.3	65.6	61.2	63.2	66.5	63.9	**65.1**
comp.sys.ibm.pc.hardware	61.3	57.1	59.1	59.8	64.7	61.9	64.8	67.3	**65.9**
comp.sys.mac.hardware	49.8	65.5	56.1	70.5	65.3	67.7	72.9	70.4	**71.5**
com.windows.x	80.4	73.2	76.5	82.4	74.7	78.3	81.9	76.6	**79.1**
misc.forsale	75.6	79.3	77.4	82.7	80.5	81.6	82.3	82.6	**82.3**
rec.autos	65.8	76.0	**70.1**	52.1	76.3	61.9	51.7	75.1	61.2
rec.motorcycles	71.1	65.9	68.3	71.8	66.7	69.1	76.5	68.7	**72.3**
rec.sport.baseball	88.9	82.5	85.5	85.8	82.1	83.8	90.1	82.7	**86.2**
rec.sport.hockey	93.7	84.1	**88.6**	90.6	83.1	86.7	89.9	86.8	88.2
sci.crypt	72.5	68.6	70.5	75.1	65.8	70.1	79.6	67.7	**73.1**
sci.electronics	58.2	57.7	57.9	57.9	60.0	58.9	62.9	60.5	**61.6**
sci.med	82.3	80.2	81.2	84.6	80.4	**82.4**	82.0	82.8	**82.4**
sci.space	76.0	73.8	**74.8**	70.2	76.3	73.1	72.7	76.4	74.5
soc.religion.christian	68.2	74.0	70.8	68.3	75.5	71.6	67.4	81.2	**73.6**
talk.politics.guns	53.2	63.6	57.9	53.5	63.9	58.1	56.4	65.5	**60.5**
talk.politics.mideast	89.3	73.9	**80.9**	87.2	73.6	79.7	83.0	74.4	78.2
talk.politics.misc	54.7	43.9	**48.6**	46.5	47.9	47.0	52.4	43.3	47.3
talk.religion.misc	30.1	5.5	8.9	32.1	13.8	18.5	29.9	18.2	**22.4**