Multi-class Text Classification using BERT-based Active Learning

Sumanth Prabhu[*]
sumanth.prabhu@swiggy.in
Applied Research, Swiggy
Bangalore

Moosa Mohamed[*]
moosa.mohamed@swiggy.in
Applied Research, Swiggy
Bangalore

Hemant Misra
hemant.misra@swiggy.in
Applied Research, Swiggy
Bangalore

ABSTRACT

Text classification finds interesting applications in the pickup and delivery services industry where customers require one or more items to be picked up from a location and delivered to a certain destination. Classifying these customer transactions into multiple categories helps understand the market needs for different customer segments. Each transaction is accompanied by a text description provided by the customer to describe the products being picked up and delivered which can be used to classify the transaction. BERT-based models have proven to perform well in Natural Language Understanding. However, the product descriptions provided by the customers tend to be short, incoherent and code-mixed (Hindi-English) text which demands fine-tuning of such models with manually labelled data to achieve high accuracy. Collecting this labelled data can prove to be expensive. In this paper, we explore Active Learning strategies to label transaction descriptions cost effectively while using BERT to train a transaction classification model. On TREC-6, AG's News Corpus and an internal dataset, we benchmark the performance of BERT across different Active Learning strategies in Multi-Class Text Classification.

CCS CONCEPTS

• **Computing methodologies** → **Active learning settings**.

KEYWORDS

Active Learning, Text Classification, BERT

1 INTRODUCTION

Most of the focus of the machine learning community is about creating better algorithms for learning from data. However, getting useful annotated datasets can prove to be difficult. In many

---

scenarios, we may have access to a large unannotated corpus while it would be infeasible to annotate all instances [Settles 2009]. Thus, weaker forms of supervision [Liang 2005; Mekala and Shang 2020; Natarajan et al. 2013; Ratner et al. 2017] were explored to label data cost effectively. In this paper, we focus on Active Learning, a popular technique under Human-in-the-Loop Machine Learning [Monarch 2019]. Active Learning overcomes the labeling bottleneck by choosing a subset of instances from a pool of unlabeled instances to be labeled by the human annotators (a.k.a the oracle). The goal of identifying this subset is to achieve high accuracy while having labeled as few instances as possible.

BERT-based [Devlin et al. 2019; Sanh et al. 2020; Yinhan Liu et al. 2019] Deep Learning models have been proven to achieve state-of-the-art results on GLUE [Wang et al. 2018], RACE [Lai et al. 2017] and SQuAD [Rajpurkar et al. 2018, 2016]. Active Learning has been successfully integrated with various Deep Learning approaches [Gal and Ghahramani 2016; Gissin and Shalev-Shwartz 2019]. However, the use of Active Learning with BERT based models for multi class text classification has not been studied extensively.

In this paper, we consider a text classification use-case in industry specific to pickup and delivery services where customers make use of short text to describe the products to be picked up from a certain location and dropped at a target location. Table 1 shows a few examples of the descriptions used by our customers to describe their transactions. Customers tend to use short incoherent and code-mixed (using more than one language in the same message[1]) textual descriptions of the products for describing them in the transaction. These descriptions, if mapped to a fixed set of possible categories, help assist critical business decisions such as demographic driven prioritization of categories, launch of new product categories and so on. Furthermore, a transaction may comprise of multiple products which adds to the complexity of the task. In this work, we focus on a multi-class classification of transactions, where a single majority category drives the transaction.

We explored supervised category classification of transactions which required labelled data for training. Our experiments with different approaches revealed that BERT-based models performed really well for the task. The train data used in this paper was labeled manually by subject matter experts. However, this was proving to be a very expensive exercise, and hence, necessitated exploration of cost effective strategies to collect manually labelled training data.

The key contributions of our work are as follows

- **Active Learning with incoherent and code-mixed data:** An Active Learning framework to reduce the cost of labelling

---

[1]In our case, Hindi written in roman case is mixed with English

Transaction Description	Category
"Get me dahi 1.5kg" **Translation :** Get me 1.5 kilograms of curd	Grocery
"Pick up 1 yellow coloured dress"	Clothes
"3 plate chole bhature"	Food
"2254/- pay krke samaan uthana hai" **Translation :** Pay 2254/- and pick up the package	Package

Table 1: Instances of actual transaction descriptions used by our customers along with their corresponding categories.

data for multi class text classification by 85% in an industry use-case.
- **BERT for Active Learning in multi-class text Classification** The first work, to the best of our knowledge, to explore and compare multiple advanced strategies in Active Learning like Discriminative Active Learning using BERT for multi-class text classification on publicly available TREC-6 and AG's News Corpus benchmark datasets.

2 RELATED WORK

Active Learning has been widely studied and applied in a variety of tasks including classification [Novak et al. 2006; Tong and Koller 2002], structured output learning [Haffari and Sarkar 2009; Roth and Small 2006; Tomanek and Hahn 2009], clustering [Bodó et al. 2011] and so on.

2.1 Traditional Active Learning

Popular traditional Active Learning approaches include sampling based on model uncertainty using measures such as entropy [Lewis and Gale 1994; Zhu et al. 2008], margin sampling [Scheffer et al. 2001] and least model confidence [Culotta and McCallum 2005; Settles 2009; Settles and Craven 2008]. Researchers also explored sampling based on diversity of instances [Dasgupta and Hsu 2008; McCallum and Nigam 1998; Settles and Craven 2008; Wei et al. 2015; Xu et al. 2016] to prevent selection of redundant instances for labelling. [Baram et al. 2004; Hsu and Lin 2015; Settles et al. 2008] explored hybrid strategies combining the advantages of uncertainty and diversity based sampling. Our work focusses on extending such strategies to Active Learning in a Deep Learning setting.

2.2 Active Learning for Deep Learning

Deep Learning in an Active Learning setting can be challenging due to their tendency of rarely being uncertain during inference and requirement of a relatively large amount of training data. However, Bayesian Deep Learning based approaches [Gal and Ghahramani 2016; Gal et al. 2017; Kirsch et al. 2019] were leveraged to demonstrate high performance in an Active Learning setting for image classification. Similar to traditional Active Learning approaches, researchers explored diversity based sampling [Sener and Savarese 2018] and hybrid sampling [Zhdanov 2019] to address the drawbacks of pure model uncertainty based approaches. Moreover, researchers also explored approaches based on Expected Model

Change [Huang et al. 2016; Settles et al. 2007; Yoo and Kweon 2019; Zhang et al. 2017] where the selected instances are expected to result in the greatest change to the current model parameter estimates when their labels are provided. [Gissin and Shalev-Shwartz 2019] proposed "Discriminative Active Learning (DAL)" where Active Learning in Image Classification was posed as a binary classification task where instances to label were chosen such that the set of labeled instances and the set of unlabeled instances became indistinguishable. However, there is minimal literature in applying these strategies to Natural Language Understanding (NLU). Our work focusses on extending such Deep Active Learning strategies for Text Classification using BERT-based models.

2.3 Deep Active Learning for Text Classification in NLP

[Siddhant and Lipton 2018] conducted an empirical study of Active Learning in NLP to observe that Bayesian Active Learning outperformed classical uncertainty sampling across all settings. However, they did not explore the performances of BERT-based models. [Prabhu et al. 2019; Zhang 2019] explored Active Learning strategies extensible to BERT-based models. [Prabhu et al. 2019] conducted an empirical study to demonstrate that active set selection using the posterior entropy of deep models like FastText.zip (FTZ) is robust to sampling biases and to various algorithmic choices such as BERT. [Zhang 2019] applied an ensemble of Active Learning strategies to BERT for the task of intent classification. However, neither of these works account for comparison against advanced strategies like Discriminative Active Learning. [Ein-Dor et al. 2020] explored Active Learning with multiple strategies using BERT based models for binary text classification tasks. Our work is very closely related to this work with the difference that we focus on experiments with multi-class text classification.

3 METHODOLOGY

We considered multiple strategies in Active Learning to understand the impact of the performance using BERT based models. We attempted to reproduce settings similar to the study conducted by Ein-Dor et al. [Ein-Dor et al. 2020] for binary text classification.

- **Random** We sampled instances at random in each iteration while leveraging them to retrain the model.
- **Uncertainty (Entropy)** Instances selected to retrain the model had the highest entropy in model predictions.
- **Expected Gradient Length (EGL)** [Huang et al. 2016] This strategy picked instances that are expected to have the largest gradient norm over all possible labelings of the instances.
- **Deep Bayesian Active Learning (DBAL)** [Gal et al. 2017] Instances were selected to improve the uncertainty measures of BERT with Monte Carlo dropout using the max-entropy acquisition function.
- **Core-set** [Sener and Savarese 2018] selected instances that best cover the dataset in the learned representation space using farthest-first traversal algorithm
- **Discriminative Active Learning (DAL)** [Gissin and Shalev-Shwartz 2019] This strategy aimed to select instances that make the labeled set of instances indistinguishable from the unlabeled pool.
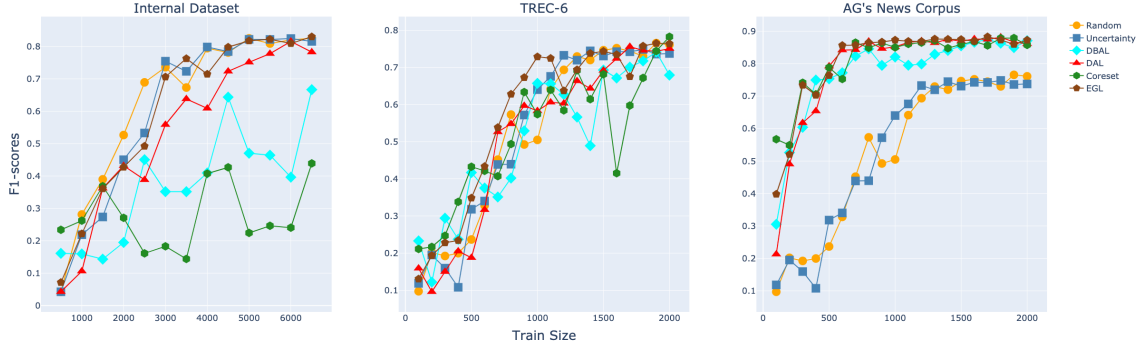
Figure 1: Comparison of F1 scores of various Active Learning strategies on the Internal Dataset(BatchSize=500, Iterations=13), TREC-6 Dataset(BatchSize=100, Iterations=20) and AG's News Corpus(BatchSize=100, Iterations=20)

4 EXPERIMENTS

4.1 Data

Our **Internal Dataset** comprised of 55,038 customer transactions each of which had an associated transaction description provided by the customer. The instances were manually annotated by a team of SMEs and mapped to one of 10 pre-defined categories. This data set was split into 44,030 training samples and 11,008 validation samples. The list of categories considered are as follows: {'Food', 'Grocery', 'Package', 'Medicines', 'Household Items', 'Cigarettes', 'Clothes', 'Electronics', 'Keys', 'Documents/Books' }
To verify the reproducibility of our observations, we considered the **TREC dataset** [Li and Roth 2002] which comprised of 5,452 training samples and 500 validation samples with 6 classes and also considered the **AG's News Corpus** [Zhang et al. 2015] which comprised of 120,000 training samples and 7,600 validation samples with 4 classes.

4.2 Experiment Setting

We leveraged DistilBERT [Sanh et al. 2020] for the purpose of the experiments in this paper. We ran each strategy for two random seed settings on the TREC-6 Dataset and AG's News Corpus and one random seed setting for our Internal Dataset. The results presented in the paper consist of 558 fine-tuning experiments (78 for our Internal dataset and 240 each for the TREC-6 dataset and AG's News Corpus). The experiments were run on AWS Sagemaker's p3 2x large Spot Instances. The implementation was based on the code[2] made available by [Gissin and Shalev-Shwartz 2019]. For the our Internal Dataset, we set the batch size to 500 and number of iterations to 13. For TREC-6 and AG's News Corpus, we set the batch size and number of iterations to 100 and 20 respectively. In each iteration, a batch of samples were identified and the model was retrained for 1 epoch with a learning rate of 3×10^{-5}. We retrained the models from scratch in each iteration to prevent overfitting [Hu et al. 2019; Siddhant and Lipton 2018].

5 RESULTS

We report results for the Active Learning Strategies using F1-score as the classification metric. Figure 1 visualizes the performance of

---

different approaches. We observe that the best F1-score that Active Learning using BERT achieves on the Internal Dataset is 0.83. A fully supervised setting with DistilBERT using the entire training data also achieved an F1-score of 0.83. Thus, we reduce the labelling cost by 85% while achieving similar F1-scores. However, we observe that no single strategy significantly outperforms the other. [Lowell et al. 2019] who studied Active Learning for text classification and sequence tagging in non-BERT models, and demonstrated the brittleness and inconsistency of Active Learning results. [Ein-Dor et al. 2020] also observed inconsistency of results using BERT in binary classification. Moreover, we observe that the performance improvements over Random Sampling strategy is also inconsistent. Concretely, on the Internal Dataset and TREC-6, we observe that DBAL and Core-set are perform poorly when compared to Random Sampling. However, on the AG's News Corpus, we observe a contradicting behaviour where both DBAL and Core-set outperform the Random Sampling strategy. Uncertainty sampling strategy performs very similar to the Random Sampling strategy. EGL performs consistently well across all three datasets. The DAL strategy, proven to have worked well in image classification, is consistent across the datasets. However, it does not outperform the EGL strategy.

6 ANALYSIS

In this section, we analyze the different AL strategies considered in the paper to understand their relative advantages and disadvantages. Concretely, we consider the following metrics -

- **Diversity** The instances chosen in each iteration need to be as different from each other as possible. We rely on the Diversity measure proposed by [Ein-Dor et al. 2020] based on the euclidean distance between the [CLS] represetations of the samples.
- **Representativeness** The AL strategies that select highly diverse instances can still have a tendency to select outlier instances that are not representative of the overall data distribution. Following [Ein-Dor et al. 2020], we capture the representativeness of the selected instances using the KNN-density measure. We quantify the density of an instance as one over the average distance between the selected instances and their K nearest neighbors within the unlabelled pool, based on the [CLS] representations.

- **Class Bias** Inspired by [Prabhu et al. 2019], we consider quantifying the level of disproportion in the class labels for each of the the AL strategies. Concretely, we measure potential Class Bias using Label Entropy [Prabhu et al. 2019] based Kullback-Leibler (KL) divergence between the ground-truth label distribution and the distribution obtained from the chosen instances.
- **Runtime** Another important factor for comparison of different AL strategies is the time taken to execute each iteration. We compare the efficiency of the AL strategies based on runtimes.

Table 2 shows the diversity and representativeness of different strategies for the datasets. We observed that Random achieved the highest scores while DAL and EGL achieved relatively higher scores compared to the remaining strategies. Core-set, which was designed to achieve high diversity, also achieves high scores in Diversity for both datasets. However, we observe that it displays inconsistent behaviour in selecting representative instances which could potentially be attributed to its tendency to select outliers. Uncertainty

Dataset	Approach	Diversity	Representativeness
TREC-6	Random	0.28	0.37
	Uncertainty	0.25	0.34
	Core-set	0.25	0.31
	DAL	0.26	0.35
	DBAL	0.25	0.33
	EGL	0.27	0.35
AG's News Corpus	Random	0.26	0.43
	Uncertainty	0.21	0.37
	Core-set	0.24	0.39
	DAL	0.25	0.4
	DBAL	0.23	0.39
	EGL	0.24	0.39

Table 2: Diversity and Representativeness for the TREC-6 Dataset and AG's News Corpus

Dataset	Approach	∩ Q	∩ S
TREC-6	Random	1.79 ± 0.01	1.79
	Uncertainty	1.26 ± 0.23	1.76
	Core-set	1.64 ± 0.15	1.77
	DAL	1.75 ± 0.02	1.79
	DBAL	1.38 ± 0.09	1.75
	EGL	1.75 ± 0.02	1.79
AG's News Corpus	Random	1.37 ± 0.01	1.39
	Uncertainty	0.87 ± 0.43	1.36
	Core-set	0.98 ± 0.42	1.22
	DAL	1.33 ± 0.03	1.38
	DBAL	0.78 ± 0.31	1.38
	EGL	1.18 ± 0.05	1.35

Table 3: Label Entropies for the TREC-6 Dataset and AG's News Corpus. ∩ Q denotes averaging across queries of a single run, ∩ S denotes the label entropy of the final collected samples

Approach	Runtime (seconds)
Random	<1
Uncertainty	84
Core-set	89
DAL	378
DBAL	120
EGL	484

Table 4: Runtimes (in seconds) for the TREC-6 Dataset per iteration for different AL strategies, with 5,000 samples
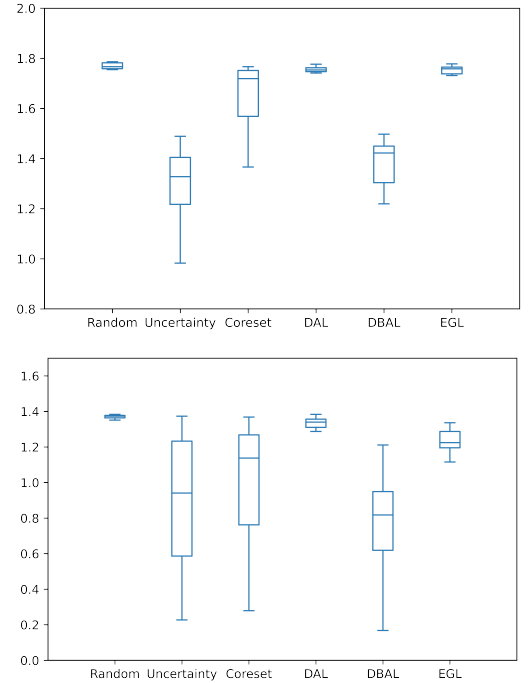

Figure 2: Class Bias scores per run for TREC-6 (top) and AG's News (bottom) across different AL strategies

achieves the lowest score in diversity and is also inconsistent in selecting representative instances. Interpreting the scores would require a deeper analysis of these observations which we leave to future work.

Table 3 shows the class bias with different approaches. Figure 2 shows the corresponding boxplots for the AL strategies. Wilcoxon signed-rank test [Rey and Neuhäuser 2011] showed that Random, DAL and EGL have significantly higher label entropies, thus, lower class bias compared to the remaining strategies for the datasets considered in the paper. However, DAL and EGL can be relatively time consuming. Table 4 shows the runtimes for each AL approach. We observed that Random is the fastest strategy while EGL is the slowest owing to its dependency on the gradient calculation.

7 CONCLUSION

In this paper, we explored multiple Active Learning strategies using BERT. Our goal was to understand if BERT-based models can prove effective in an Active Learning setting for multi-class text classification. We observed that EGL performed reasonably well across datasets in multi-class text classification. Moreover, Random, EGL and DAL captured diverse and representative samples with relatively lower class bias. However, unlike Random, EGL and DAL had longer execution times. In future work, we plan to perform a deeper analysis of the observations and also explore ensembling approaches by combining the advantages of each strategy to explore potential performance improvements.

REFERENCES

Yoram Baram, Ran El-Yaniv, and Kobi Luz. 2004. Online Choice of Active Learning Algorithms. *J. Mach. Learn. Res.* 5 (Dec. 2004), 255–291.

Zalán Bodó, Zsolt Minier, and Lehel Csató. 2011. Active Learning with Clustering. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010 (Proceedings of Machine Learning Research, Vol. 16)*, Isabelle Guyon, Gavin Cawley, Gideon Dror, Vincent Lemaire, and Alexander Statnikov (Eds.). JMLR Workshop and Conference Proceedings, Sardinia, Italy, 127–139. http://proceedings.mlr.press/v16/bodo11a.html

Aron Culotta and Andrew McCallum. 2005. Reducing Labeling Effort for Structured Prediction Tasks. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2* (Pittsburgh, Pennsylvania) *(AAAI'05)*. AAAI Press, 746–751.

Sanjoy Dasgupta and Daniel Hsu. 2008. Hierarchical Sampling for Active Learning *(ICML'08)*. Association for Computing Machinery, New York, NY, USA, 208–215. https://doi.org/10.1145/1390156.1390183

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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

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)*. Association for Computational Linguistics, Online, 7949–7962. https://doi.org/10.18653/v1/2020.emnlp-main.638

Yarin Gal and Zoubin Ghahramani. 2016. Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference. arXiv:1506.02158 [stat.ML]

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian Active Learning with Image Data. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 1183–1192. http://proceedings.mlr.press/v70/gal17a.html

Daniel Gissin and Shai Shalev-Shwartz. 2019. Discriminative Active Learning. arXiv:1907.06347 [cs.LG]

Gholamreza Haffari and Anoop Sarkar. 2009. Active Learning for Multilingual Statistical Machine Translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, Suntec, Singapore, 181–189. https://www.aclweb.org/anthology/P09-1021

Wei-Ning Hsu and Hsuan-Tien Lin. 2015. Active Learning by Learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (Austin, Texas) *(AAAI'15)*. AAAI Press, 2659–2665.

Peiyun Hu, Zachary C. Lipton, Anima Anandkumar, and Deva Ramanan. 2019. Active Learning with Partial Feedback. arXiv:1802.07427 [cs.LG]

Jiaji Huang, Rewon Child, Vinay Rao, Hairong Liu, Sanjeev Satheesh, and Adam Coates. 2016. Active Learning for Speech Recognition: the Power of Gradients. arXiv:1612.03226 [cs.CL]

Andreas Kirsch, Joost R. van Amersfoort, and Y. Gal. 2019. BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. In *NeurIPS*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. *arXiv preprint arXiv:1704.04683* (2017).

David D. Lewis and William A. Gale. 1994. A Sequential Algorithm for Training Text Classifiers. *CoRR* abs/cmp-lg/9407020 (1994). arXiv:cmp-lg/9407020 http://arxiv.org/abs/cmp-lg/9407020

Xin Li and Dan Roth. 2002. Learning Question Classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1* (Taipei, Taiwan)

(COLING '02). Association for Computational Linguistics, USA, 1–7. https://doi.org/10.3115/1072228.1072378

Percy Liang. 2005. Semi-supervised learning for natural language. In *MASTER'S THESIS, MIT*.

David Lowell, Zachary Chase Lipton, and Byron C. Wallace. 2019. Practical Obstacles to Deploying Active Learning. In *EMNLP/IJCNLP*.

Andrew McCallum and Kamal Nigam. 1998. Employing EM and Pool-Based Active Learning for Text Classification. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 350–358.

Dheeraj Mekala and Jingbo Shang. 2020. Contextualized Weak Supervision for Text Classification. In *Association for Computational Linguistics (ACL'20)*.

Robert (Munro) Monarch. 2019. *Human-in-the-Loop Machine Learning*. Manning Publications.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. 2013. Learning with Noisy Labels. In *Neural Information Processing Systems (NIPS'13)*.

Blaž Novak, Dunja Mladenič, and Marko Grobelnik. 2006. Text Classification with Active Learning. In *From Data and Information Analysis to Knowledge Engineering*, Myra Spiliopoulou, Rudolf Kruse, Christian Borgelt, Andreas Nürnberger, and Wolfgang Gaul (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 398–405.

Ameya Prabhu, Charles Dognin, and Maneesh Singh. 2019. Sampling Bias in Deep Active Classification: An Empirical Study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4058–4068. https://doi.org/10.18653/v1/D19-1417

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Association for Computational Linguistics*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing*.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid Training Data Creation with Weak Supervision. In *Proceedings of the VLDB Endowment (VLDB'17)*.

Denise Rey and Markus Neuhäuser. 2011. *Wilcoxon-Signed-Rank Test*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1658–1659. https://doi.org/10.1007/978-3-642-04898-2_616

Dan Roth and Kevin Small. 2006. Margin-Based Active Learning for Structured Output Spaces. In *Machine Learning: ECML 2006*, Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 413–424.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL]

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active Hidden Markov Models for Information Extraction.

Ozan Sener and Silvio Savarese. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. arXiv:1708.00489 [stat.ML]

Burr Settles. 2009. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison.

Burr Settles and Mark Craven. 2008. An Analysis of Active Learning Strategies for Sequence Labeling Tasks *(EMNLP '08)*. Association for Computational Linguistics, USA, 1070–1079.

Burr Settles, Mark Craven, and Soumya Ray. 2007. Multiple-Instance Active Learning. In *Proceedings of the 20th International Conference on Neural Information Processing Systems* (Vancouver, British Columbia, Canada) *(NIPS'07)*. Curran Associates Inc., Red Hook, NY, USA, 1289–1296.

Burr Settles, Mark Craven, and Soumya Ray. 2008. Multiple-Instance Active Learning. In *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis (Eds.), Vol. 20. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2007/file/a1519de5b5d44b31a01de013b9b51a80-Paper.pdf

Aditya Siddhant and Zachary C. Lipton. 2018. Deep Bayesian Active Learning for Natural Language Processing: Results of a Large-Scale Empirical Study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2904–2909. https://doi.org/10.18653/v1/D18-1318

Katrin Tomanek and Udo Hahn. 2009. Reducing Class Imbalance during Active Learning for Named Entity Annotation. In *Proceedings of the Fifth International Conference on Knowledge Capture* (Redondo Beach, California, USA) *(K-CAP '09)*. Association for Computing Machinery, New York, NY, USA, 105–112. https://doi.org/10.1145/1597735.1597754

Simon Tong and Daphne Koller. 2002. Support Vector Machine Active Learning with Applications to Text Classification. *J. Mach. Learn. Res.* 2 (March 2002), 45–66. https://doi.org/10.1162/153244302760185243

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Empirical Methods in Natural Language Processing*.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. 2015. Submodularity in Data Subset Selection and Active Learning. In *Proceedings of the 32nd International Conference on International*

Conference on Machine Learning - Volume 37 (Lille, France) *(ICML'15)*. JMLR.org, 1954–1963.

Zuobing Xu, Ram Akella, and Yi Zhang. 2016. Incorporating Diversity and Density in Active Learning for Relevance Feedback.

Myle Ott Yinhan Liu, 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. *Computing Research Repository* arXiv:1907.11692 (2019). https://arxiv.org/pdf/1907.11692.pdf version 1.

Donggeun Yoo and In So Kweon. 2019. Learning Loss for Active Learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 93–102. https://doi.org/10.1109/CVPR.2019.00018

L. Zhang. 2019. An Ensemble Deep Active Learning Method for Intent Classification. *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence* (2019).

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-Level Convolutional Networks for Text Classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1* (Montreal, Canada) *(NIPS'15)*. MIT Press, Cambridge, MA, USA, 649–657.

Ye Zhang, Matthew Lease, and Byron C. Wallace. 2017. Active Discriminative Text Representation Learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, California, USA) *(AAAI'17)*. AAAI Press, 3386–3392.

Fedor Zhdanov. 2019. Diverse mini-batch Active Learning. arXiv:1901.05954 [cs.LG] https://arxiv.org/pdf/1901.05954.pdf

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 - Volume 1* (Manchester, United Kingdom) *(COLING'08)*. Association for Computational Linguistics, USA, 1137–1144.