Active Learning for Text Classification and Fake News Detection

Marko Sahan
Dept. of Computer Science, FEE, CTU in Prague
Karlovo náměstí 13, 121 35, Prague
+420224357337
sahanmar@fel.cvut.cz

Vaclav Smidl
Dept. of Computer Science, FEE, CTU in Prague
Karlovo náměstí 13, 121 35, Prague
+420224357337
smidlva1@fel.cvut.cz

Radek Marik
Dept.of Telecommunication Engineering, FEE, CTU in Prague
Technicka 2, 166 27, Prague
+420224354058
radek.marik@fel.cvut.cz

ABSTRACT

Supervised classification of texts relies on the availability of reliable class labels for the training data. However, the process of collecting data labels can be complex and costly. A standard procedure is to add labels sequentially by querying an annotator until reaching satisfactory performance. Active learning is a process of selecting unlabeled data records for which the knowledge of the label would bring the highest discriminability of the dataset. In this paper, we provide a comparative study of various active learning strategies for different embeddings of the text on various datasets. We focus on Bayesian active learning methods that are used due to their ability to represent the uncertainty of the classification procedure. We compare three types of uncertainty representation: i) SGLD, ii) Dropout, and iii) deep ensembles. The latter two methods in cold- and warm-start versions. The texts were embedded using Fast Text, LASER, and RoBERTa encoding techniques. The methods are tested on two types of datasets, text categorization (Kaggle News Category and Twitter Sentiment140 dataset) and fake news detection (Kaggle Fake News and Fake News Detection datasets). We show that the conventional dropout Monte Carlo approach provides good results for the majority of the tasks. The ensemble methods provide more accurate representation of uncertainty that allows to keep the pace of learning of a complicated problem for the growing number of requests, outperforming the dropout in the long run. However, for the majority of the datasets the active strategy using Dropout MC and Deep Ensembles achieved almost perfect performance even for a very low number of requests. The best results were obtained for the most recent embeddings RoBERTa

Keywords

Active learning, deep learning ensembles, fake news classification, Fast Text, LASER, natural language processing, RoBERTa, text classification, uncertainty representation.

1. INTRODUCTION

The development of a text classifier on a new problem requires the availability of the training data and their labels. Labeling involves human annotators and a common practice is to label as many text documents as possible, train a classifier and search for new data and labels if the performance is unsatisfactory. Random choice of the documents for the data set extension can be costly because the new documents may not bring new information for the classification. Active learning strategy aims to select among available unlabeled documents those that the classifier is most uncertain about and queries an annotator for their labels. Therefore, it has the potential to greatly reduce the effort needed for the development of a new system. While it was introduced almost two decades ago, recent improvements in deep learning motivate our attempt to revisit the topic. For example, SVM-based active learning approaches for text classification date back to 2001 [22], where the superiority of active learning over random sampling was demonstrated. Since deep recurrent and convolutional neural networks achieve better classification results, Bayesian active learning methods for deep networks gained popularity especially in image classification [10], [15].

The Bayesian approach is concerned with querying labels for data for which the classifier predicts the greatest uncertainty. The uncertainty is quantified using the so-called acquisition function, such as predictive variance or predictive entropy [18]. While different acquisition functions often provide similar results, different representations of predictive distribution yield much more diverse results. The most popular approach using Dropout MC [10] has been tested on text classification [1] and named entity recognition [19], [15], however other techniques such as Langevin dynamics [26] and deep ensembles [13] are available. Deep ensembles often achieve better performance [3], [21] but require higher computational cost since they train an ensemble of networks after each extension of the data set. One potential solution of this problem has been recently proposed in [23], where the ensemble is not trained from a fresh random initialization after each query but initialized randomly around the position of the ensembles from the previous iteration. In this contribution, we test this approach and compare it with the dropout MC and Langevin dynamics representations. We also provide sensitivity study for the choice of the hyperparameters.

Active learning for fake news detection was considered in [4] using uncertainty based on probability of classification. It was later extended to a context aware approach [5]. An entropy based approach has been presented in [12] using an ensemble of three different kinds of networks.

2. Methods

Throughout the paper, we will use three different embedding algorithms such as Fast Text [16], LASER [2] and RoBERTa [14]. RoBERTa is a modified BERT transformer model [6] based on multi-head attention layers [24]. Transformer models provide state of the art results in context understanding and it is advantageous to compare behavior of active learning algorithms with respect to different embedding techniques. Representation of the i-th text document x_i is calculated as the mean value from sentence embeddings of all sentences in the text

$$\boldsymbol{x}_i = \frac{1}{|D_i|}\sum_{j\in D_i} f_{Sentence\ embed}(C^{(j)}),$$

where D_i is the set of vectors where each vector represents a sentence in the i-th document, $|D_i|$ is a cardinality of D_i, $C^{(j)}$ is a

matrix of j-th sentence where words are encoded with one hot or byte pair encoding [20] technique and $f_{Sentence\ embed}$ is a function that creates sentence embeddings with respect to the given one-hot or byte pair encoded words. Fast Text encoding is made with respect to one-hot encoded words. LASER and transformer based models take byte-per encoded words as an input. Sentence embeddings for LASER and RoBERTa are output of deep neural network models. Sentence embedding for the Fast Text model is calculated as a mean value of all embeddings of words in the sentence.

For supervised classification, each document embedding x_i has to have an associated label y_i. We are concerned with binary classification for simplicity, however, an extension to multiclass is straightforward. We assume that for the full corpus of text documents $X = [x_1, \ldots, x_n]$, only an initial set of $l_0 \ll n$ labels is available, $Y^{(0)} = [y_0, \ldots, y_{l_0}]$, splitting the full set X to the labeled, $X^{(0)} = [x_1, \ldots, x_{l_0}]$, and unlabeled parts, $X \backslash X^{(0)}$. Active learning is defined as a sequential extension of the training data set. In each iteration, $l = 1, \ldots, L$, the algorithm computes entropy of the predictive probability distribution for each document in the unlabeled dataset and selects the index of the document with the highest entropy (entropy acquisition function), formally:

$$k_l = arg\ \max_{k \in K} E_{p(\theta|X^{(l-1)}, Y^{(l-1)})}(H(y_k|\theta)) \qquad (1)$$

$$H(y_k|\theta) = E_{p(y_k|\theta, x_k)}(-\log p(y_k|\theta, x_k)) \qquad (2)$$

where K is the set of indexes of all unlabeled documents, E is the expectation operator over the posterior probability of the classifier parameters θ trained on all labeled data $p(\theta|X^{(l-1)}, Y^{(l-1)})$ and $H(y_k|\theta)$ is conditional entropy of the predicted class for x_k. The document of the selected index is sent to the human annotator with a *request* for labeling. When the selected text is annotated, the text is added with its label to the labeled data set $X^{(l)} = [X^{(l-1)}, x_{k_l}]$, $Y^{(l)} = [Y^{(l-1)}, y_{k_l}]$.

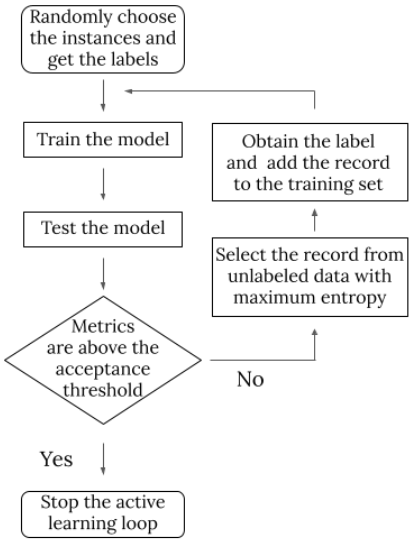

Figure 1: Flowchart of the active learning algorithm

The procedure is repeated L times. The active learning process is visualised in figure 1.

The key component of the method is a representation of the posterior distribution of the parameter θ. Due to the complexity of the neural networks it is always represented by samples, with a different method of their generation. We will compare the following methods: i) SGLD: Stochastic Gradient with Langevin dynamics [26], which adds additional noise to the gradient in stochastic gradient descent, ii) Dropout MC: is an extension of the ordinary dropout that samples binary mask multiplying output of a layer, hence stopping propagation through all neurons where zeros is sampled through the network. The extension applies the sampled mask even for predictions generating samples from the predictive distribution [10], iii) Deep ensembles: consist of N networks trained in parallel from different initial conditions [13], and iv) Softmax uncertainty: is the most simple approach that uses only one network to estimate a single value $\hat{\theta}$ and maximizing entropy $H(y_k|\hat{\theta})$ instead of the expectation (1) [4]. Ensembles based approach is the current state-of-the-art in active learning [3].

While many of these have been tested in active learning, the majority of authors assumed that after each step of active learning, the network training starts from the initial conditions. This is clearly suboptimal, since the information from the previous training is lost. A simple solution was presented in [23], where it was argued that estimated results from the previous step can be used as centroids around which the new initial point is sampled. Since this is a form of a warm-start, we also test warm-start strategies for Dropout. The methods for representation of parametric uncertainty are:

Deep Ensemble Filter (DEnFi):

is a deep ensemble method with 10 neural networks in the ensembles and warm-start training strategy [13] using weights of the ensemble members in the previous iteration as initial conditions for the new ensemble.

Algorithm 1: DEnFi training algorithm

```
for network in ensemble:
    if not initialized:
        mean = 0
        variance = q_0
        epochs = initialization_epochs
    else:
        mean = nn.weights
        variance = q
        epochs = warm_start_epochs
network.weights = gaussian_noise(mean, variance)
network.train(data, epochs)
```

Algorithm 2: DEnFi predicting algorithm

```
results = empty_array()
for network in ensemble:
    results.append(network.predict(data_point))
```

Each weight is perturbed by an additive Gaussian noise of variance q which is a hyperparameter. In our experiments, the ensemble is

trained with parameters initialization_epochs = 2000 on the initial data and with additional warm_start_epochs = 700 epochs after each extension of the learning data set.

Dropout MC:

is the standard algorithm [10] that trains only a single network with sampled dropout indices and uses the sampling even in the prediction step. Generation of the Monte Carlo prediction is obtained by sampling different values of the dropout binary variable and one forward pass of the network for each sample. We study three versions of the algorithm: i) cold-start with 3000 epochs after each request, ii) warm-start with weights from the previous iteration perturbed by an additive noise of variance q with 700 epochs and iii) hot-start, with 50 epochs after each request without perturbation (hot start is a warm-start method with q=0). Dropout rate is 0.5.

SGLD:

variance of the noise added to the gradient descent:

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \frac{|X^{(l)}|}{|X^{(l)}|_{minibatch}} \frac{\eta_n}{2} (\nabla L(X^{(l)}, Y^{(l)}, \hat{\theta}_n)) + \epsilon_n,\ \epsilon_n \sim N(0, \eta_n I)$$

where L is a loss function and ϵ_n is noise with covariate matrix $\eta_n I$ where η_n is the learning rate of the SGD algorithm. We choose initial value of η_1 to be 0.01 and update rule $\eta_{n+1} = \frac{\eta_n}{b-3000} + 0.05$. The noise ϵ_n is added to the gradient only after 3000 of initial training epochs. Then we draw 50 samples with 100 epochs between consecutive samples to avoid correlation.

Softmax uncertainty:

The simplest approach to uncertainty representation is a single neural network with a softmax output layer that considers uncertainty as the output of the softmax score. We add it to comparison since it has been applied to active learning in [4]. The model is trained to run 2000 epochs on the initial data with additional 200 epochs after each extension of the learning data set.

3. Experiments

The methods were compared on the positive/negative tweets from the Tweets Dataset [11], 5 pairs of categories from the News Category Dataset [17] and two types of the news datasets for fake news detection [7], [8]. Documents from the News Category dataset were downloaded using links provided in the dataset. The names of the tested categories are shown in table 3, figure 2 and in figure 3.

Specifically, we compared active learning and random sampling strategies for different settings of document embeddings and different representations of uncertainty. Each experiment was initiated by random choice of the initial training set of $l_0 = 10$ samples from 1000 text documents (500 text documents per category), which were the initial 1000 documents of the datasets. For each experiment $L = 200$ requests for annotation are simulated. The document selection follows the ϵ -greedy approach [25], i.e. the sample with maximum acquisition function (1) is accepted with probability $\epsilon = \frac{exp(l-40)}{exp(l-40)+1}$. A random document is selected for labeling if not accepted. After each request, the classification performance is evaluated on the remaining part of the selected dataset (i.e. on the 990 text documents in the first evaluation) using the area under the ROC curve (AUC) metrics [9]. In order to make the results statistically valid, we repeat the described simulation loop 10 times for Twitter and News Category datasets and 5 times for Fake News and Fake News Detection datasets.

3.1 Hyperparameter Tuning

The classification network was designed as a feed-forward NN with one dense layer of 100 neurons with sigmoid activation functions, and softmax output layer. Warm start versions of both Dropout MC and DEnFi have hyperparameter q that governs the perturbation of the previous result before training on the extended dataset. Tuning of this hyperparameter was performed by a grid search for both the Fast Text and RoBERTa text encoding techniques. Since the main focus of the paper is on the effect of the warm-start strategy, the results of the effect of the noise variance q for DEnFi and Dropout MC is displayed in table 1 and table 2 for active learning strategy on the Tech vs Science task and Fake News Detection dataset, respectively.

Based on table 1 it is clearly seen that the variance of the perturbation noise related to the best result is increasing with the number of requests. We conjecture that the variance has the role of a selector of the exploration/exploitation tradeoff. Low variance favors exploitation and improves quickly, higher variance implies less accurate guesses in the initial iterations but better performance in the long run. The results from table 1 indicate that higher values of q are not performing well thus the range of q was reduced for tuning of the hyperparameter for the fake news datasets. The increase of the optimum value of the noise variance with the number of requests is also visible in table 2, especially for the RoBERTa encoded data. Since calibration of the variance for all methods and all datasets would be too computationally expensive, we ran all remaining experiments with $q = 0.3$ for Fast Text encoding and $q = 0.1$ for RoBERTa and LASER encodings. However, tuning of this hyperparameter for an application scenario or an adaptive tuning strategy offers clearly a potential for further improvement.

3.2 Text Category Classification

A comparison of AUC of the active learning strategy after 200 requests for all tested algorithms and Fast Text encoding with respect to Tweets and News Category datasets is reported in table 3. This experiment was designed for comparison of various uncertainty representations. While all methods except for the Softmax Uncertainty achieved best results on some datasets, the most consistent results were delivered by the DEnFi and Dropout MC methods. The SGLD performance results exhibited the highest variance within the 10 runs of the method. The results for three pair categories and an active learning strategy are visualized in figure 4. Since it is the least reliable method, we will not study it any further. It is clear from table 3 and figure 4 that the simple Softmax Uncertainty provides the worst results. Since Dropout MC has comparable computational complexity, the use of simple Softmax Uncertainty is always suboptimal.

Detailed analysis of the DEnFi and Dropout MC strategies is provided in figure 2 for Tweets and News Category datasets. For better insight, we also display the performance of the random

sampling learning strategy where training data are extended using randomly sampled documents. Note that for random sampling, Dropout MC outperforms consistently DEnFi on all tasks. We conjecture that this is due to the robustness of the dropout regularization. However, the power of DEnFi becomes apparent with the increasing number of requests. It is improving slower than Dropout MC at the beginning, but improves faster, thus outperforming Dropout MC in the long run. We conjecture that this is due to better exploration capability of the DEnFi while Dropout MC excels at exploitation. The speed of improvement depends on the complexity of the learning task. For simpler tasks (such as crime vs. Good News), AUC over 0.98 is achieved quickly. However, for more complex tasks, such as Positive vs Negative Tweets, the number of data needed for improvement is much higher. The active learning is on par with the random sampling up to 125 requests and even after 200 requests, the AUC is below 0.7 indicating poor performance. Note that the active learning strategy of DEnFi starts improving over the random sampling sooner than Dropout MC with a sharper slope which indicates a high probability of obtaining the same profile as the other datasets in the long run.

3.3 Fake News Classification

The fake news data experiment is made under the same initial conditions as in Section 3.2 but for a wider choice of the document encoding techniques. Based on results from Section 3.2 we show comparison of only the DEnFi and Dropout warm-start for three different types of encodings (Fast Text, LASER, RoBERTa) on the two fake news datasets in figure 3. The results for fake news datasets are similar to those for the Twitter and News Category datasets. Advantage of the DEnFi approach is significant only for the Fake News Detection dataset with Laser embeddings. For other cases the difference is negligible. Under the exploration/exploitation hypothesis, this means that the data sets are well separable and more sophisticated exploration capabilities of DEnFi are not needed.

The best performance of the active learning strategy was achieved for the RoBERTa embedding. Note that the embedding has much greater influence on the speed of learning than uncertainty representation.

4. Conclusion

We have studied the suitability of various uncertainty representations for the task of active text classification. The established Dropout MC methodology was compared against deep ensembles, SGLD and simple softmax strategy. To reduce the computational cost, we studied warm-start strategies for both ensembles (called DEnFi) and Dropout MC, that resulted in significant reduction of training epochs per the active learning iteration. The resulting methods exhibit a different tradeoff between exploration and exploitation. While Dropout MC has been found to be more reliable in random sampling strategy and improving faster at the beginning of active learning, the DEnFi was found to prefer exploration sacrificing performance at the beginning but outperforming Dropout MC in the longer run of active learning. The highest deviation of the random strategies between DEnFi and Dropout MC resulted in 0.04 AUC in favor of Dropout MC. However, for the active learning strategy both Dropout MC and DEnFi converged to the same numbers.

Sensitivity of the active learning to the choice of embedding was evaluated on the fake news datasets. It was observed that the most recent embedding method (RoBERTa) facilitated the fastest learning and that the choice of embedding was more important than the uncertainty representation. However, methods of active learning achieved significantly faster learning than the random sampling approach (the difference between the random and the active learning strategies differs from 0.015 AUC up to 0.080 AUC) on all tested datasets of various complexity.

ACKNOWLEDGMENTS

The authors acknowledge the support of the OP VVV MEYS funded project CZ.02.1.01/0.0/0.0/16_019/0000765 “Research Center for Informatics”

REFERENCES

[1] Bang An, Wenjun Wu, and Huimin Han. 2018. Deep active learning for text classification. In *Proceedings of the 2nd International Conference on Vision, Image and Signal Processing*, pages 1–6.

[2] Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

[3] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. 2018. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9368–9377.

[4] Sreyasee Das Bhattacharjee, Ashit Talukder, and Bala Venkatram Balantrapu. 2017. Active learning based news veracity detection with feature weighting and deep-shallow fusion. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 556–565. IEEE.

[5] Sreyasee Das Bhattacharjee, William J Tolone, and Ved Suhas Paranjape. 2019. Identifying malicious social media contents using multi-view context-aware active learning. *Future Generation Computer Systems*, 100:365–379.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[7] 2018a. Fake news dataset. https://www.kaggle.com/c/fake-news, accessed on 20 february 2021.

[8] 2018b. Fake news detection dataset. https://www.kaggle.com/jruvika/fake-news-detection, accessed on 20 february 2021.

[9] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874.

[10] 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-Volume 70*, pages 1183–1192. JMLR. org.

[11] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford. http://help.sentiment140.com/for-students/ accessed on 26 june 2020.

[12] Md Saqib Hasan, Rukshar Alam, and Muhammad Abdullah Adnan. 2020. Truth or lie: Pre-emptive detection of fake news in different languages through entropy-based active learning and multi-model neural ensemble.

[13] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402– 6413.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

[15] David Lowell, Zachary C Lipton, and Byron C Wallace. 2019. Practical obstacles to deploying active learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 21–30.

[16] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[17] Rishabh Misra. 2018. News category dataset, 06. https://www.kaggle.com/rmisra/ news-category-dataset, accessed on 26 june 2020.

[18] Claude E Shannon. 1948. A mathematical theory of communication. *Bell system technical journal*, 27(3):379– 423.

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

[20] Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. 1999. Byte pair encoding: A text compression scheme that accelerates pattern matching. Technical report, Citeseer.

[21] Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D Sculley, Joshua Dillon, Jie Ren, and Zachary Nado. 2019. Can you trust your model's uncertainty? evaluating predictive uncer- tainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13969–13980.

[22] Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66.

[23] Lukas Ulrych and Vaclav Smidl. 2020. Deep ensemble filter for active learning. Technical Report 2383, Institute of Information Theory and Automation.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

[25] Christopher John Cornish Hellaby Watkins. 1989. Learning from delayed rewards.

[26] Max Welling and Yee W Teh. 2011. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688.

5. Appendix

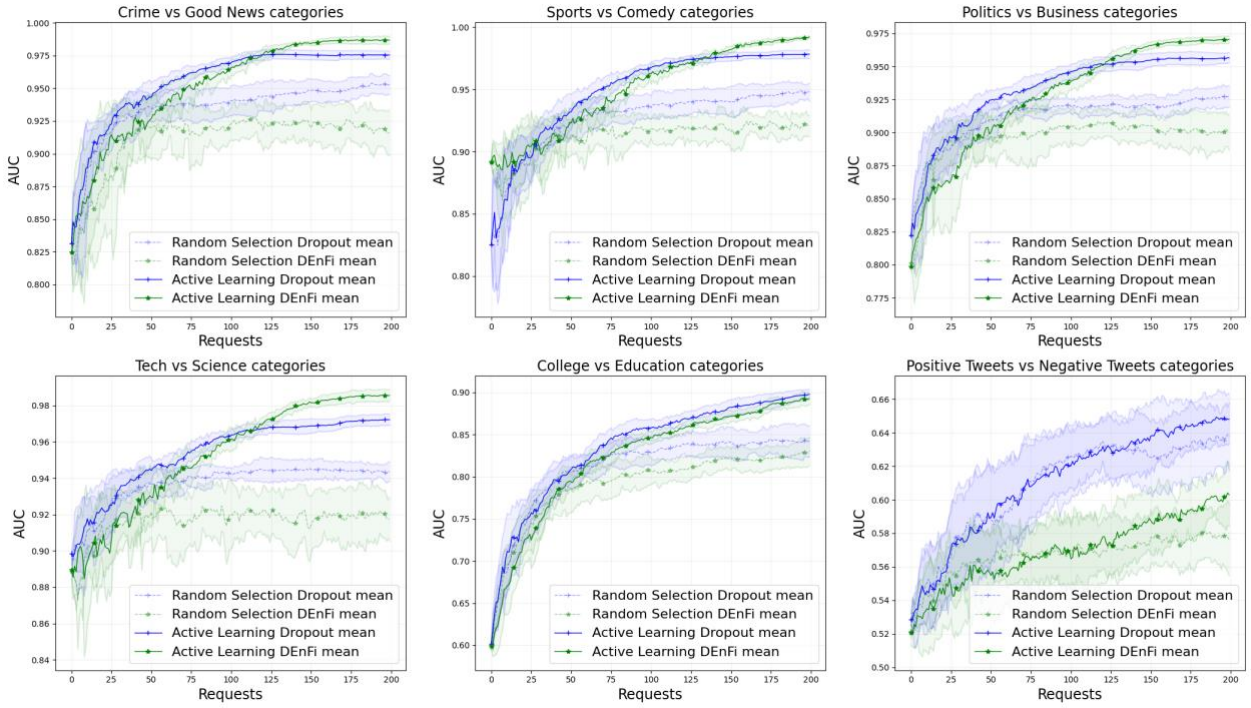

Figure 2: Evolution of the mean AUC with growing number of requests for DEnFi, Dropout MC warm-start algorithms and six pairs of categories. The uncertainty bounds are illustrated as one standard deviation from the mean value with respect to 10 runs. Both DEnFi and Dropout MC were initially trained on 10 labeled text documents before sequential learning strategies were initialized

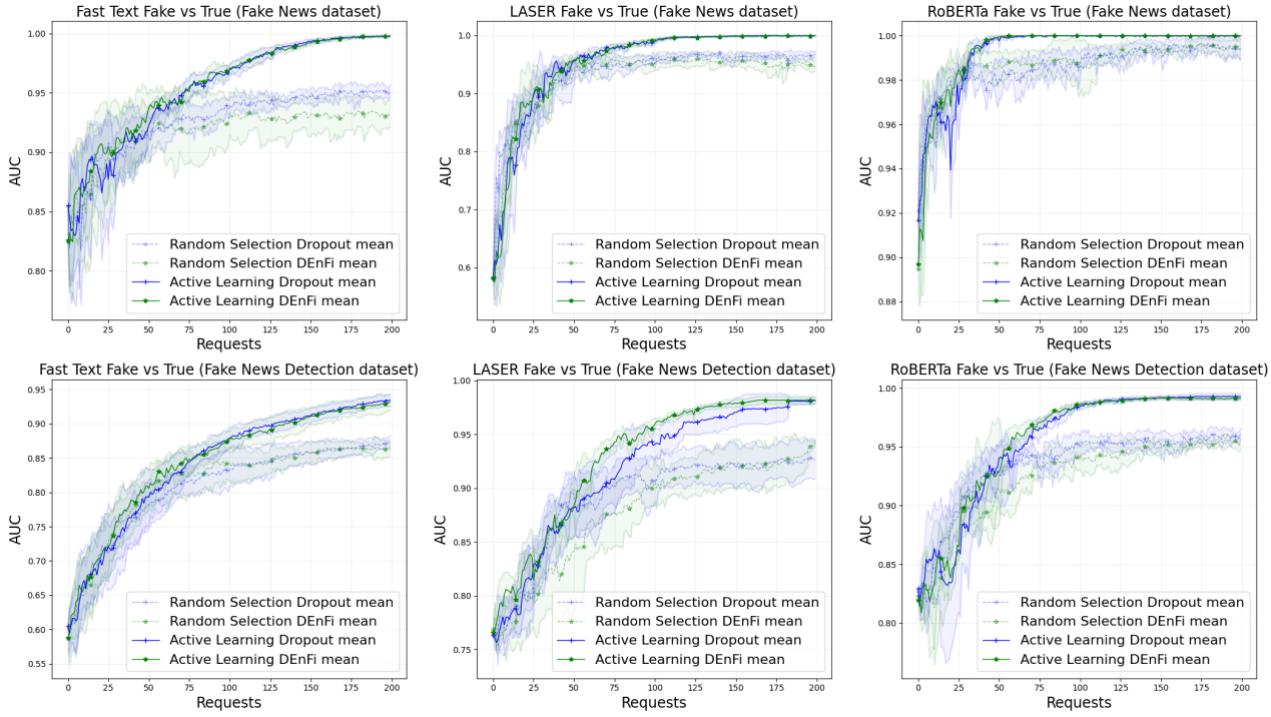

Figure 3: Evolution of the mean AUC with growing number of requests for DEnFi, Dropout MC warm-start algorithms, two datasets: Fake News (top row) and Fake News Detection (bottom row), and three embeddings: Fats Text (left column), Laser (middle column) and RoBERTa (right column). The uncertainty bounds are illustrated as one standard deviation from the mean value calculated with respect to 5 runs. Both DEnFi and Dropout MC were initially trained on 10 labeled text documents before sequential learning strategies were initialized

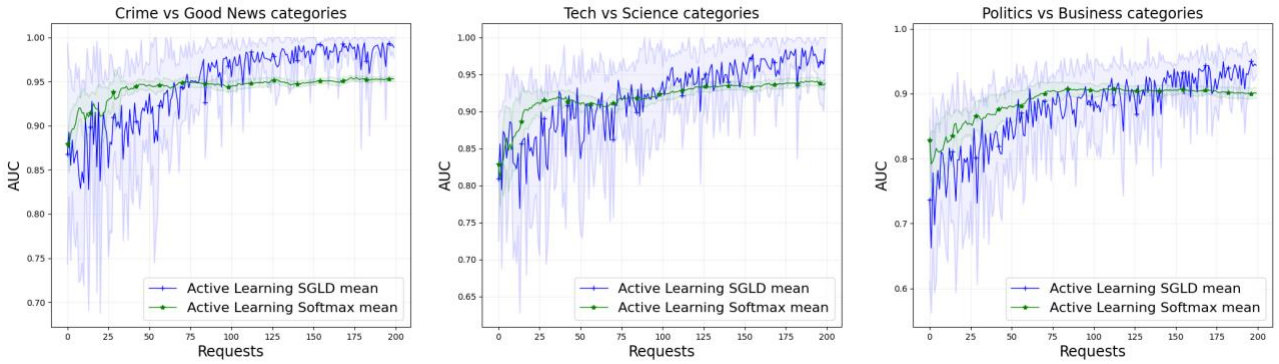

Figure 4: Evolution of the mean AUC with growing number of requests for SGLD, Softmax Uncertainty algorithms, and three pairs of categories. The uncertainty bounds are illustrated as one standard deviation from the mean value with respect to 10 runs. Both SGLD and Softmax Uncertainty were initially trained on 10 labeled text documents before sequential learning strategies were initialized

Table 1: **Fast Text** and **RoBERTa** encoding based AUC of text classification after selected number of requests of the active learning using DEnFi and Dropout MC warm-start for various selection of the perturbation noise q. Average over 10 runs in the Tech vs Science categories. The best result is denoted by a star, results within one standard deviation of the winner are displayed in bold font.

(**a**): **Fast Text** DEnFi

Noise variance	**AUC after # of iterations**			
	50	**100**	**150**	**200**
0.1	**0.945***	**0.968***	0.974	0.976
0.2	0.932	0.964	0.976	0.978
0.3	0.930	0.961	**0.982***	0.986
0.4	0.909	0.948	0.976	**0.990***
0.6	0.874	0.921	0.952	0.979
1	0.805	0.871	0.906	0.941

(**b**): **Fast Text** Dropout

Noise variance	**AUC after # of iterations**			
	50	**100**	**150**	**200**
0.1	**0.936**	**0.966***	0.972	0.976
0.2	**0.938***	**0.966***	**0.981***	0.983
0.3	0.920	0.956	**0.980**	**0.989***
0.4	0.917	0.955	0.976	**0.988**
0.6	0.894	0.948	0.972	**0.986**
1	0.859	0.914	0.941	0.970

(**c**) **RoBERTa** DEnFi

Noise variance	**AUC after # of iterations**			
	50	**100**	**150**	**200**
0.1	**0.935**	**0.970***	**0.982**	0.988
0.2	**0.940***	0.954	**0.983***	**0.990***
0.3	0.903	0.930	0.967	0.987
0.4	0.883	0.911	0.945	0.976
0.6	0.799	0.853	0.885	0.945
1	0.661	0.750	0.818	0.875

(**d**) **RoBERTa** Dropout

Noise variance	**AUC after # of iterations**			
	50	**100**	**150**	**200**
0.1	**0.930***	**0.970**	0.982	0.986
0.2	**0.923**	**0.971***	**0.986***	**0.990***
0.3	0.917	0.959	0.982	**0.990***
0.4	0.878	0.949	0.974	**0.990***
0.6	0.822	0.908	0.952	0.980
1	0.643	0.741	0.862	0.921

Table 2: **Fast Text** and **RoBERTa** encoding based AUC of text classification after selected number of requests of the active learning using DEnFi and Dropout warm-start for various selection of the perturbation noise q. Average over 5 runs on the Fake and True news categories (Fake News Detection dataset). The best result is denoted by a star, results within one standard deviation of the winner are displayed in bold font.

(**a**) **Fast Text** DEnFi

Noise variance	**AUC after # of iterations**			
	50	**100**	**150**	**200**
0.2	**0.794**	**0.886***	**0.910**	0.942*
0.3	**0.806***	**0.877**	**0.911***	**0.932**
0.4	**0.806***	0.866	**0.901**	**0.911**

(**b**): **Fast Text** Dropout warm-start

Noise variance	**AUC after # of iterations**			
	50	**100**	**150**	**200**
0.2	**0.806**	**0.879***	**0.916***	**0.949***
0.3	**0.804**	**0.884***	0.914	0.939
0.4	**0.809***	**0.878**	**0.912**	**0.944**

(**c**): **RoBERTa** DEnFi

Noise variance	**AUC after # of iterations**			
	50	**100**	**150**	**200**
0.1	**0.929***	**0.986***	**0.992***	0.991
0.2	**0.923**	0.975	0.990	**0.998***
0.3	**0.891**	0.935	0.955	0.975

(**d**): **RoBERTa** Dropout warm-start

Noise variance	**AUC after # of iterations**			
	50	**100**	**150**	**200**
0.1	**0.941***	**0.984***	**0.992***	**0.993**
0.2	0.917	0.974	**0.992***	**0.995***
0.3	0.917	0.959	0.982	0.990

Table 3: Average AUC over 10 runs for five different algorithms after 200 iterations of active learning and six different datasets. The best result is denoted by a star, results within one standard deviation of the winner are displayed in bold font.

Method	**Crime/ Good News**	**Sports/ Comedy**	**Politics/ Business**	**Tech/ Science**	**Education/ College**	**Pos./Neg. Tweets**
SGLD	**0.989***	0.968	0.944	0.984	0.881	0.621
DEnFi, $q = 0.3$	0.987	**0.992***	**0.971***	**0.986**	**0.893**	0.603
Dropout cold-start	0.975	0.978	0.957	0.972	**0.898***	**0.648**
Dropout warm-start, $q = 0.3$	0.978	0.979	0.954	0.973	0.877	**0.657***
Dropout warm-start, $q = 0$	0.978	0.951	0.944	**0.989***	0.824	0.561
Softmax Uncertainty	0.953	0.939	0.901	0.938	0.800	0.609