# Ensembles Active Learning for Text Classification

**Anonymous COLING submission**

## Abstract

Machine learning approach to classification is based on learning parameters of black box model describing relation between the recorded data samples and their class labels. The process of data labels collection for the purposes of model training can be complex and costly. Therefore, the number of data record is often much higher than the number of labels, but the labels can be obtained by querying an annotator. Active learning is a process of selection of unlabeled data records for which knowledge of the label would bring the highest discriminability of the dataset. Various methods for active learning have been proposed in many different fields that use supervised learning models. In this project, we study suitability and propose new active learning approach of a text classification problem. We compare existing state-of-the-art active learning classifiers to deep ensembles model with deeper study of uncertainty calibration for both existing and new approaches.

## 1 Introduction

Active learning strategy lets the machine learning models iteratively and strategically query the labels of some instances for reducing human labeling efforts. Our goal is to show the active learning performance on text data. People has already been solving active learning problem for anomaly detection (Das *et al.* , , 2018), image processing (Gal *et al.* , , 2017) (Sener & Savarese, , 2017), named entity recognition (Shen *et al.* , , 2017), (Lowell *et al.* , , 2019), (Burkhardt *et al.* , , 2018), etc..

Modern approach of automating the labeling process of huge amount of unlabeled data, is not optimal. People are randomly choosing unlabeled text data. These data are annotated by the subject matter experts, and used for training and testing the models. If the model performance is weak after the training, more text documents are selected and annotated. This approach is costly because nobody knows how much text documents must be selected in order to have good model scores. Our active learning strategy proposes selection of unlabeled text data that the model is not certain about. Unlabeled text data are given to a subject matter expert to provide the labels. Discussed problem was introduced almost two decades ago. Some SVM based active learning approaches for text classification date back to 2001 (Tong & Koller, , 2001), where are shown different querying strategies and results of the active learning superiority over random sampling strategies. Even though SVM show good results, deep recurrent and convolutional neural networks gained higher popularity due to their efficiency in text classification (Lowell *et al.* , , 2019), (An *et al.* , , 2018). Active learning let's us to start with lower amount of training data, and iteratively extend the dataset. In our work we are aimed on choosing the data, which our model is not certain about. One of the most popular methods for deep learning uncertainty representation is a dropout (Gal *et al.* , , 2017). Langevin dynamics (Welling & Teh, , 2011) and DEnFi also show good representation of the model's uncertainty. In this work we are extending our dataset with only one sample per active learning iteration. However, it was also shown that the strategies, which sample batches with more than one sample, also perform good results (An *et al.* , , 2018).

The project shows comparison of different active learning methods and their modification for text classification. The main point of this project is to provide comprehensive comparison between deep ensembles and dropout representation algorithms. Basing on (Snoek *et al.* , , 2019) and (Lakshminarayanan *et al.* , , 2017), it was shown that ensemble deep learning algorithms give the best performance both for text and image processing data. Wide variety of acquisition functions for unlabeled data sampling was shown and compared in (Gal *et al.* , , 2017), In this project we illustrate the results with respect to entropy based acquisition function. We believe that a comparison of dropout and ensembles based models for active learning text classification will give better understanding of the problem and will bring more clarity to their strong and weak parts.

## 2 Methods

Supervised learning is defined as learning from the data with known target value. Specifically, for the classification problem, the target value is the class, where each data point belongs.

We would like to commence our formal definition with the data. Let $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^n$ and $\mathbf{y} \in \mathcal{Y} = \{[0,1]^T, [1,0]^T\}$, where $\mathbf{x}$ is feature vector of size $n$, and $\mathbf{y}$ is its label assigned to the data instance $\mathbf{x}$ from space $\mathcal{X}$. Each value from space $\mathcal{Y}$ can be represented as a one hot representation, which is a vector consisting from ones and zeros. In the case of binary classification $\mathbf{y} \in \{[0,1]^T, [1,0]^T\}$, where the first class is represented as $\mathbf{y} = [1,0]^T$ and the second class is represented as $\mathbf{y} = [0,1]^T$. As a good example of previous definition, $\mathbf{x}$ can be a text document (represented in a mathematical form in order to meet a definition above) and $\mathbf{y}$ can be its category, such as sports or comedy. As seen from this example, the label and the text are forming a tuple. In this work we are considering our data as tuples of variables $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$.

The training set is usually a subset of all available data on which the optimization is performed, the rest of the data is used for validation (Vapnik, , 2013).

### 2.1 Embedding

Transformer models such as BERT (Devlin *et al.* , , 2018) or other different modifications showed their strength in context understanding. However, we assume that Fast Text (Mikolov *et al.* , , 2018) text encoding is efficient enough for our text classification purposes. Instance $\mathbf{x}$ is calculated as a mean value from all words embeddings in the text

$$\mathbf{x}_i = \frac{1}{|\mathcal{D}_i|} \sum_{j \in \mathcal{D}_i} f_{\text{Fast text}}(\mathbf{c}^{(j)}),$$

where $\mathcal{D}_i$ is the set of indices of all words in $i$-th document in the common vocabulary, $\mathbf{c}^{(j)}$ is $j$-th one hot encoded word vector and $f_{\text{Fast text}}$ is a function that creates Fast Text embeddings with respect to given one-hot encoded word.

### 2.2 Active Learning

As mentioned in previous sections, the training set $\tilde{\mathbf{X}} \times \tilde{\mathbf{Y}}$ is only a subset of all available data. It is important that each point $\mathbf{x}$ in the training set has a label $\mathbf{y}$. In many situations labels are expensive to obtain, and the number $N$ of all available samples $\mathbf{x} \in \mathbf{X}$ is large, while labels are available only for the initial subset $J_0 = \{1, \ldots, N_0\}, N_0 \ll N$. This problem is known as semi-supervised learning.

We consider a setup in which we can ask for a label for an arbitrary $\mathbf{x}$, that can be provided for example by a human (annotator). We assume that getting labels needs some time and is very expensive. The task is to choose which sample we will ask to label.

The active learning problem is defined as a sequence of supervised learning problems. Specifically, we assume that the initial sets for supervised learning are $\mathbf{X}_0 = \{\mathbf{x}_i\}_{i \in J_0}$ and $\mathbf{Y}_0 = \{\mathbf{y}_i\}_{i \in J_0}$. We consider a sequence of $U$ questions $u = \{1, \ldots, U\}$, in each question we select an index $j_u$ and ask to obtain the label $\mathbf{y}_{j_u}$ for data record $\mathbf{x}_{j_u}$. The index set and the data sets are extended as follows

$$J_u = \{J_{u-1}, j_u\}, \qquad \mathbf{X}_u = \{\mathbf{X}_{u-1}, \mathbf{x}_{j_u}\}, \qquad \mathbf{Y}_u = \{\mathbf{Y}_{u-1}, \mathbf{y}_{j_u}\}.$$

The task of active learning is to optimize the selection of indices $j_u$ to reach as good classification metrics with as low number of questions as possible.

The decision task for this particular problem can be written as

$$j_u^* = \underset{j_u \in J \setminus J_u}{\operatorname{argmin}} (\mathbb{E}_{\pi_u^*} L^*), \tag{1}$$

where $\mathbb{E}_{\pi_u^*} L^*$ is the expected loss that is dependent on an action given question $u$, and $J$ is the space of all indices. Character "$*$" is used only for distinguishing active learning loss from the loss function which is used for different models. Using this approach, we will be able sequentially select indices from $\mathbf{X}$ and ask for a label from $\mathbf{Y}$, that will help us to get higher scores faster than in the case of random choice of indices.

### 2.3 Bayesian neural networks

## 3 Experiments

The active learning simulation compares different algorithms with respect to two strategies: i) random, and ii) entropy based selection of text documents. In this paper we decided to use positive negative tweets

| **Entropy based acquisition function** | Crime vs Good News | Sports vs Comedy | Politics vs Business | Tech vs Science | Education vs College | Positive vs Negative Tweets |
|---|---|---|---|---|---|---|
| SGLD | **0.989** | 0.968 | 0.944 | 0.984 | 0.881 | 0.621 |
| DEnFi V1 | **0.990**$^*$ | 0.973 | 0.952 | 0.985 | **0.896** | **0.652** |
| DEnFi V2 | 0.987 | **0.992**$^*$ | **0.971**$^*$ | **0.986** | **0.893** | 0.603 |
| Dropout cold start | 0.975 | $0.978z$ | 0.957 | 0.972 | **0.898**$^*$ | **0.648** |
| Dropout hot start | 0.978 | 0.979 | 0.954 | 0.973 | 0.877 | **0.657**$^*$ |
| Dropout hot start w noise | 0.978 | 0.951 | 0.944 | **0.989**$^*$ | 0.824 | 0.561 |

$(*)$ maximal value per column

Table 1: Binary classification mean AUC results over 10 runs for 6 different algorithms with respect to 200 active learning iterations and six different pairs of categories. Bold values represent intersection of a mean value with respect to one standard deviation interval from a maximal value $(*)$

from Tweets Dataset (Go A., , 2009) and 5 pairs of categories form News Category Dataset (Misra, , 2018). Names of the categories are shown in table 1 and in figure 1. The categories were chosen with respect to different classification complexity. We randomly choose initial training set that has 10 samples. Described 10 random samples are chosen from 1000 text documents (500 text documents per category). We reduce the size of all above mentioned datasets to 1000 documents. Each dataset is split on testing and training data.

Next we initialize two runs. First run is random selection of the text documents and second run is entropy based selection of the documents. Both runs start with the same training dataset, and then they chose additional training documents based on their strategies. We consider continuous new data selection from 1000 documents dataset and imitating annotators labeling process. All in all, we repeat text samples selection 200 times ($U = 200$). We select 10 random samples in the beginning, train our model, and make a prediction on the complement to the training dataset (990 text documents). As a further step, we select 1 new sample from the set on which the prediction was done. New text sample for labeling is chosen with respect to acquisition function or random choice. Before we extend our training dataset with a new labeled sample, we calculate the area under ROC curve (AUC) metrics on the complement to the dataset (990 text documents). This process is repeated 200 times. Thus, by the end of the simulation, our training set will have 210 text documents and testing set (complement to a training set) will have 790 data samples. In order to make our results statistically valid, we repeat the described simulation loop 10 times.

### 3.1 Neural Network Approaches Comparison

In table 1 is shown comparison of different neural networks based active learning algorithms with respect to AUC mean. The results in table 1 are shown for 200 active learning iterations and entropy acquisition function. We belief that DEnFi V2 and Dropout cold start algorithms show the best performance with respect to shown results.

In figure 1 is illustrated whole evolution of AUC mean over 10 runs with uncertainty bounds for six pairs of categories. It is seen that in four out of six plots DEnFi V2 shows better active learning results. It is also seen that DenFi mimic behavior of Dropout curve for College vs Education categories. In comparison to the categories where DEnFi showed its superiority over Dropout we can see that AUC values were higher than 95%, whereas maximal AUC value for College vs Education is around 90%. It can be said that DEnFi V2 did not have enough active learning iterations in order to explore the dataset better. It is seen that in the final steps of a simulation, DEnFi V2 curve is converging to Dropout curve. The behavior that was seen for College vs Education categories can be also observed for Tweets Dataset. The AUC scores are very low and it is seen that both DEnFi V2 and Dropout are copying random selection strategy evolution. However, we can observe that in the right part of the plot, DEnFi V2 active learning strategy starts to show better scores than random selection strategy. In this case we can also conclude that DEnFi did not have enough training samples for overcoming Dropout.

### 3.2 Hot Start Noise Calibration

Both DEnFi V2 and Dropout with hot start are quite sensitive with respect to added noise. DEnFi V2 and Dropout hot start mean AUC results with respect to different gaussian noise variance configuration for Tech vs Science categories are shown in table 2. It is clearly seen that for little amount of training data strategy with less noise works better. However, when the amount of training data is increasing, the strategy with higher amount of noise is more efficient. Described trend is observed both for DEnFi V2

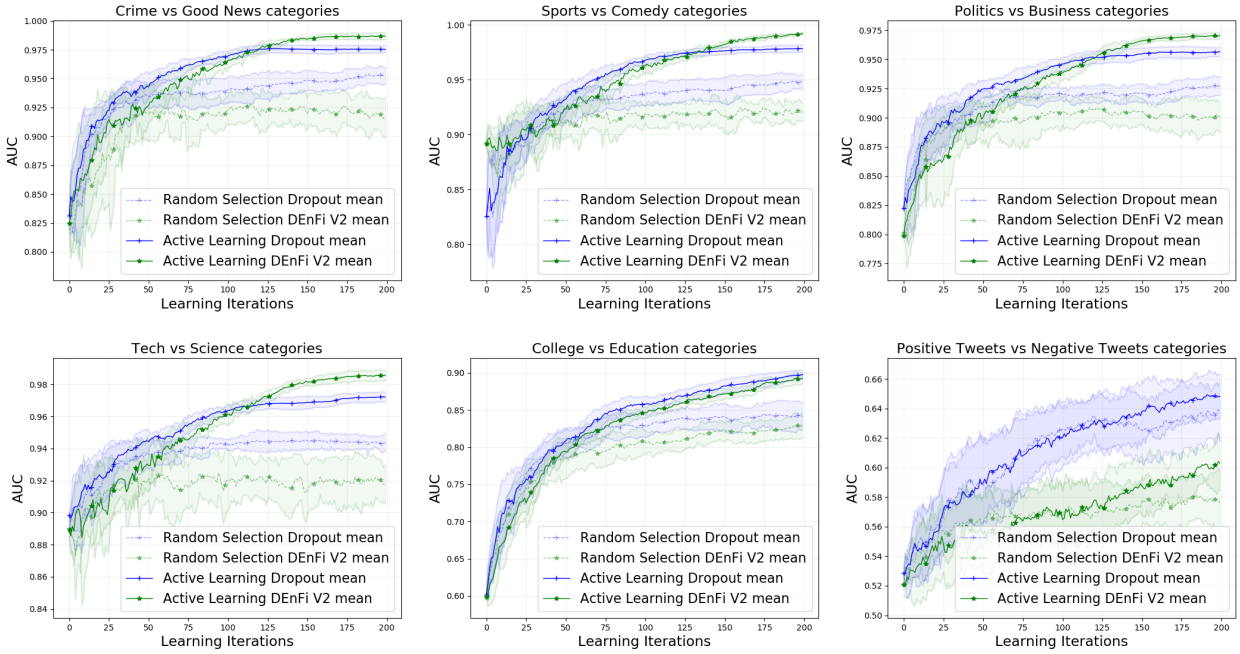


Figure 1: AUC mean evolution with respect to learning iterations for DEnFi V2, Dropout cold start algorithms and six pairs of categories. The uncertainty bounds are illustrated as one standard deviation from the mean value. Both DEnFi V2 and Dropout were initially trained on 10 labeled text documents before sequential learning strategies were initialized

and Dropout algorithms.

## 4 Conclusion

In this work it was illustrated and verified that ensembles exploration power can overcome dropout based active learning algorithm if querying process lasts long enough. Described behavior was perfectly seen both on the plots for News Category Dataset where the classification scores were high and for Twitter Dataset where the classification scores were low. Another important aspect of the work are hot start methods and noise calibration. DEnFi V2 method showed great results by using the knowledge from previous iterations. The trend of increasing noise and better classification for higher amount of training documents may let us to reach higher scores with adaptive noise addition. Combination of the facts mentioned above can be a strong basis for the future research.

| **Noise** | **Active learning iterations** | | | | |
|---|---|---|---|---|---|
| **variance** | 0 | 50 | 100 | 150 | 200 |
| 0.1 | 0.867 | **0.945**$^*$ | **0.968**$^*$ | 0.974 | 0.976 |
| 0.2 | **0.893**$^*$ | 0.932 | 0.964 | 0.976 | 0.978 |
| 0.3 | 0.890 | 0.930 | 0.961 | **0.982**$^*$ | 0.986 |
| 0.4 | 0.886 | 0.909 | 0.948 | 0.976 | **0.990**$^*$ |
| 0.6 | 0.846 | 0.874 | 0.921 | 0.952 | 0.979 |
| 1 | 0.777 | 0.805 | 0.871 | 0.906 | 0.941 |

($*$) maximal value per column

(a) DEnFi V2

| **Noise** | **Active learning iterations** | | | | |
|---|---|---|---|---|---|
| **variance** | 0 | 50 | 100 | 150 | 200 |
| 0.1 | **0.902**$^*$ | **0.936** | **0.966**$^*$ | 0.972 | 0.976 |
| 0.2 | **0.901** | **0.938**$^*$ | **0.966**$^*$ | **0.981**$^*$ | 0.983 |
| 0.3 | 0.899 | 0.920 | 0.956 | **0.980** | **0.989**$^*$ |
| 0.4 | 0.900 | 0.917 | 0.955 | 0.976 | **0.988** |
| 0.6 | 0.898 | 0.894 | 0.948 | 0.972 | **0.986** |
| 1 | 0.866 | 0.859 | 0.914 | 0.941 | 0.970 |

($*$) maximal value per column

(b) Dropout hot start with noise

Table 2: DENFI V2 and Dropout hot start with noise binary classification mean AUC results over 10 runs for Tech vs Science categories with respect to different noise variance configuration and number of active learning iterations. Bold values represent intersection of a mean value with respect to one standard deviation interval from a maximal value ($*$)

## References


An, Bang, Wu, Wenjun, & Han, Huimin. 2018. Deep Active Learning for Text Classification. *Pages 1–6 of: Proceedings of the 2nd International Conference on Vision, Image and Signal Processing.*

Burkhardt, Sophie, Siekiera, Julia, & Kramer, Stefan. 2018. Semisupervised bayesian active learning for text classification. *In: Bayesian Deep Learning Workshop at NeurIPS.*

Das, Shubhomoy, Islam, Md Rakibul, Jayakodi, Nitthilan Kannappan, & Doppa, Janardhan Rao. 2018. Active anomaly detection via ensembles. *arXiv preprint arXiv:1809.06477.*

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

Gal, Yarin, Islam, Riashat, & Ghahramani, Zoubin. 2017. Deep bayesian active learning with image data. *Pages 1183–1192 of: Proceedings of the 34th International Conference on Machine Learning-Volume 70.* JMLR. org.

Go A., Bhayani R., Huang L. 2009. *Twitter sentiment classification using distant supervision.*

Lakshminarayanan, Balaji, Pritzel, Alexander, & Blundell, Charles. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Pages 6402–6413 of: Advances in neural information processing systems.*

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

Mikolov, Tomas, Grave, Edouard, Bojanowski, Piotr, Puhrsch, Christian, & Joulin, Armand. 2018. Advances in Pre-Training Distributed Word Representations. *In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

Misra, Rishabh. 2018 (06). *News Category Dataset.*

Sener, Ozan, & Savarese, Silvio. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489.*

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

Snoek, Jasper, Ovadia, Yaniv, Fertig, Emily, Lakshminarayanan, Balaji, Nowozin, Sebastian, Sculley, D, Dillon, Joshua, Ren, Jie, & Nado, Zachary. 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Pages 13969–13980 of: Advances in Neural Information Processing Systems.*

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

Vapnik, Vladimir. 2013. *The nature of statistical learning theory.* Springer science & business media.

Welling, Max, & Teh, Yee W. 2011. Bayesian learning via stochastic gradient Langevin dynamics. *Pages 681–688 of: Proceedings of the 28th International Conference on Machine Learning (ICML-11).*