Incorporating Active learning Methods with Contextualized Sentence Representations

Uri Katz Danielle Hausler

Abstract

Active learning (AL) is a commonly used technique that aims to reduce the annotation cost while maximizing the predictive performance of a model. This technique is also useful in many natural language processing (NLP) tasks such as classification, named entity recognition and word sense disambiguation. Only few works have examined the influence of text representation on active learning. We have investigated the behavior of several active learning methods combining representation and uncertainty sampling with two different contextual text representation methods. The experimental results show a relation between the representation type and the AL performance, based on the representation ability to separate classes.

1 Introduction

Many modern machine learning tasks require large amount of labelled training data. Labelled data is often costly , sometimes requires annotation by experts which is time consuming or just hard to get. Active learning methods aim to overcome this problem by sampling the most informative instances and getting the same performance with less data.

In the context of NLP, active learning techniques have shown to be helpful in many text classification tasks. Shen et al. (2004) incorporate informativeness, representativeness and diversity measurements to actively choose the best instances for named entity recognition task. Zhu et al. (2008) combined uncertainty and density measurements to optimize the selected instances for word sense disambiguation and text classification tasks. Li et al. (2012b) reduced the annotation cost for imbalanced sentiment classification by two complementary classifiers to select the most informative minority class samples and majority class samples.

Text classification tasks involve transformation of the text into numeric representations that the classifying algorithm can process. In the recent years,the idea of dynamic word embeddings became popular. those embeddings enable different representations to the same word, based on the context that the word is surrounded by. Such contextualized word representation like BERT and ELMO have created by using deep neural language models (Devlin et al., 2018; Peters et al., 2018) . Bert is based on a deeper model than ELMO and has a deep bidirectional transformer for encoding with self attention layers. BERT has reached the state-of-the-art in many NLP benchmarks, including sentence classification and semantic textual similarity. However extracting sentences similarity by BERT from a large corpus might require many hours of computation. One possible solution to handle this issue is to map each sentence to a fix sized vector, by averaging BERT embbedings (the last output layer of BERT). A recent alternative solution suggested by Reimers and Gurevych (2019) is to fine-tune the BERT model to produce semantically meaningful sentence representations. As shown in their paper , this method was superior to Average BERT embeddings, effiecient to compute and achieved state-of-the-art performance on the SentEval toolkit.

Only few works have examined the influence of text representation on active learning for text classification.The representative power of sentence BERT embeddings haven't been tested yet in the domain of active learning. In this work we would like to combine different contextual text representation methods - Average BERT embeddings and Sentence BERT embeddings, with active learning strategies in order to answer the following questions:

1. Will AL strategies that incorporate representation measurements be more efficient?
2. What is the influence of the representation on AL? Will similarity-based representation improve AL performance in sentiment classification tasks?

2 Related works

Lu et al. (2019) compared different word representations in Active Learning for text classification. They used classic representation methods such as BOW , Latent Dirichlet Allocation and modern representation methods such as FastText and BERT. However, their work contained only one query strategy that incorporated uncertainty and representative sampling. We aim to deepen their analysis by examining several new query strategies that accounts for both information and representation measurements. In addition, due to the superior performance of BERT we will focus only on BERT representations. We will use two different sentence representations based on averaging or fine tuning BERT embeddings and we will look for interactions between the query strategy and the representation. There are also some similar studies that considered the representation measurement in active learning (Zhu et al., 2008; Huang et al., 2010; Figueroa et al., 2012). Recent work by Zhdanov (2019) introduced an efficient query method that incorporates the informativeness and diverse of the unlabelled samples. Since their method was superior to many other baselines including uncertainty sample, we will try to reproduce their method and results on other representations.

3 Methodology

3.1 Query strategies

Query strategies can be divided into two groups: model-naïve and model-dependent.The former strategies are based on the text representation solely while the latter based on the model probability estimation for $P(y_u|x_u)$ or any other scoring system. Here we will define the query strategies building blocks used in the experiments.

Algorithm 1: The Active learning setup

```
Input: a pool of labeled samples ($X_L$) ,
  unlabeled samples ($X_U$) , and a querying
  strategy $\pi$
initialization
- select $n$ random samples from $X_U$
- send samples to oracle for labeling and
  moved to $X_L$
- train classifier $f(X_L)$
for iteration=1 ... $k$ do
    - querying strategy $\pi(x_U, x_L, f)$ calculate
      score for every $x_U \in X_U$.
    - top score $n$ samples sent to the oracle for
      labeling and move to $X_L$.
    - train classifier $f(X_L)$
end
*not all querying strategies utilize $x_L$ and $f$
```

3.1.1 Model-naïve query strategies

The model-naïve strategies are aiming to query samples by properties of their vector representations. First, we define the metric of relation between two vectors as follow:

$$sim(\mathbf{x_i}, \mathbf{x_j}) = \frac{\mathbf{x_i} \cdot \mathbf{x_j}}{\|\mathbf{x_i}\| \|\mathbf{x_j}\|} \tag{1}$$

Then we can define the query strategies :

Max Diversity-Representative:

The Max Diversity-Representative (MDR) takes the most Representative samples and from those it samples the most diverse ones. Let samples $x_i, x_j \in X_U$ then the representativeness of x_i is defined by:

$$R(x_i) = \frac{1}{|U|} \sum_{j=1}^{|U|} sim(x_i, x_j) \tag{2}$$

Which is the average similarity between a sample and all the other samples in the unlabeled pool.

Let sample $x_i \in X_U$ and $x_j \in X_L$, The diversity of x_i is defined by:

$$D(x_i) = \frac{1}{|L|} \sum_{j=1}^{|L|} 1 - sim(x_i, x_j) \tag{3}$$

Which is the average of dissimilarity between a sample from the unlabeled pool and all the labeled samples. By querying diverse samples we are aiming to span the contextual space better.

Combining the last two methods (equations (2),(3)), we get for $x_i \in X_U$:

$$MDR(x_i) = D(x_i) \cdot R(x_i) \tag{4}$$

MDR strategy is aiming to increase the diversity of the training with texts that represent large groups of samples.

3.1.2 Model-dependent query strategies

Model dependent query strategies utilize the model's uncertainty in the inference phase, Also known as uncertainty sampling.

Least Confidence:
The classifier's uncertainty when dealing with a sample can be assessed in a probabilistic manner using least confidence method. The decisions boundary of a probabilistic binary model is the boundary that defines the sample's class. Samples with $P(y|x_i)$ close to the decisions boundary are a sign to uncertainty , i.e the model is not sure to which class the sample is belong to. Least Confidence (LC) of $x_i \in X_U$ for both binary and multi-class models is defined by:

$$LC(x_i) = 1 - \underset{y}{\operatorname{argmax}}\, P(y|x_i) \tag{5}$$

High LC corresponds to $P(y|x_i)$ centered around the decision boundary which means higher uncertainty. In our experiment we used SVM model which is not a probabilistic model ,therefore we used adjusted variation of LC that is defined as :

$$LC(x_i) = \underset{dj}{\operatorname{argmin}}\, |d_j| \tag{6}$$

where $j \in (1, 2, ...k)$ of k classes and d is the distance from the decision boundary. Again, as before ,being close to the decision boundary means uncertainty regarding the classification.

Kmeans clustering:
Let $T \subseteq X_U$ be a subset of samples from the unlabeled pool. Unsupervised Kmeans algorithm define k clusters over T. Using euclidean distance on the $\ell 2$-normalized vectors is proportional to cosine similarity. Therefore we can query samples which are far from (Diverse) or close to (Representativeness) the defined cluster's centroid.
In our Experiment we implement two different approaches combined with the Least confidence uncertainty strategy (see next). :

Diverse mini-Batch Active Learning (DBAL):
Here we implement an approach suggested in (Zhdanov, 2019). First we measure the uncertainty over X_U using Least confidence strategy. Then we take the βn most uncertain samples ,in our experiment β=5 and n is the number of samples the strategy takes in each iteration.We cluster those βn examples to n clusters and select n different samples closest to the cluster centers. By doing so, the strategy is aiming to select a diverse set of samples with better ability to span the training space.

Diversity kmeans (DIV-Kmeans):
Here we combine Least confidence and Kmeans as well , but with only two clusters. Then we select the samples with the max distance from the centroids. The clusters here are large groups of similar samples and by taking the maximal distance samples we are increasing the diversity of the training set.
Query By Committee:
The Query By Committee (QBC) strategy involves a committee C = $\{f_1$,...,$f_C\}$ of classifying models which are trained on the labeled pool X_L. Each member of the committee predicts the labels of X_U and votes for them. The most uncertain sample is the sample that the committee most disagree on. For a sample x$_i$ *let* $V(y_i,_j)$ be the number of votes for label j, for $j \in (1, 2.., k)$ of k classes.
then we can define the sample's voting disagreement by the voting entropy (VE):

$$VE(x_i) = -\sum_j \frac{V(y_i,_j)}{|C|} log \frac{V(y_i,_j)}{|C|} \tag{7}$$

This strategy aims at choosing the samples with the max VE thus:

$$QBC(x_*) = \underset{x}{\operatorname{argmax}}\, VE(x) \tag{8}$$

QBC strategy is based on uncertainty sampling and might be susceptible to outliers. In order to improve the traditional QBC strategy we added a density measurement to choose the most informative (uncertain) samples and representative in terms of density. We chose a k-nearest neighbour approach to evaluate the density of an unlabeled sample. for a set of K nearest neighbours: $\mathrm{KNN}(x_i) = \{x_1 \dots x_k\}$, the sample's KNN density is:

$$DENS(x_i) = \frac{\sum_{x_j \in KNN(x_i)} \cos(x_i, x_j)}{K} \tag{9}$$

We define QBC-KNN (QN) as:

$$QN(x_i) = QBC(x_i) \cdot DENS(x_i) \tag{10}$$

4 Datasets

MR:
Movie reviews snippets, with 5,331 positive snippets and 5,331 negative snippets (Pang and Lee, 2005).

TREC:
Dataset for question classification , with 5,000 labeled questions and six different classes (Li and Roth, 2002). This dataset is commonly used in text classification papers (Sun et al., 2019; Howard and Ruder, 2018; Kim, 2014) .

TOXIC:
The dataset from Jigsaw/Google Toxic Comment Classification Challenge on Kaggle. Contains Wikipedia comments which have been labeled for toxic behavior. In our experiment we sampled 5,000 comments,with the same class ratio.Since the dataset is class imbalanced (1:10 proportion) we used it to assess performance over imbalance training.

4.1 Dataset Pre-Process

We transformed the sentences into two different vector representations, by Bert Sentence transformers [1] with 'bert-large-nli-stsb-mean-tokens' model and by bert-as-a-service-repository [2] we averaged BERT sentence embeddings. As a result each sentence was represented as a 1X1024 vector (by BERT Sentence transformers) or 1X768 vector (by bert-as-a-service).

Dataset	c	N	Representation	F1
MR	2	10,661	AVGBert	0.8
			SenBert	0.81
Toxic	2	5,000	AVGBert	0.79
			SenBert	0.78
TREC	6	5,000	AVGBert	0.84
			SenBert	0.71

Table 1: Summary statistics for the datasetss. c: Number of target classes. N: number of samples.Representation: The representation used in the experiment. F1: average F1 over 5 folds CV over the entire dataset

5 Experiments

5.1 classifying models

We avoid using deep learning classifiers which requires a large amount of data to outperform other approaches (Kowsari et al., 2019) ,instead we used a Support Vector Machine (SVM) (Joachims, 1998) classifier.

5.2 Experimental Setup

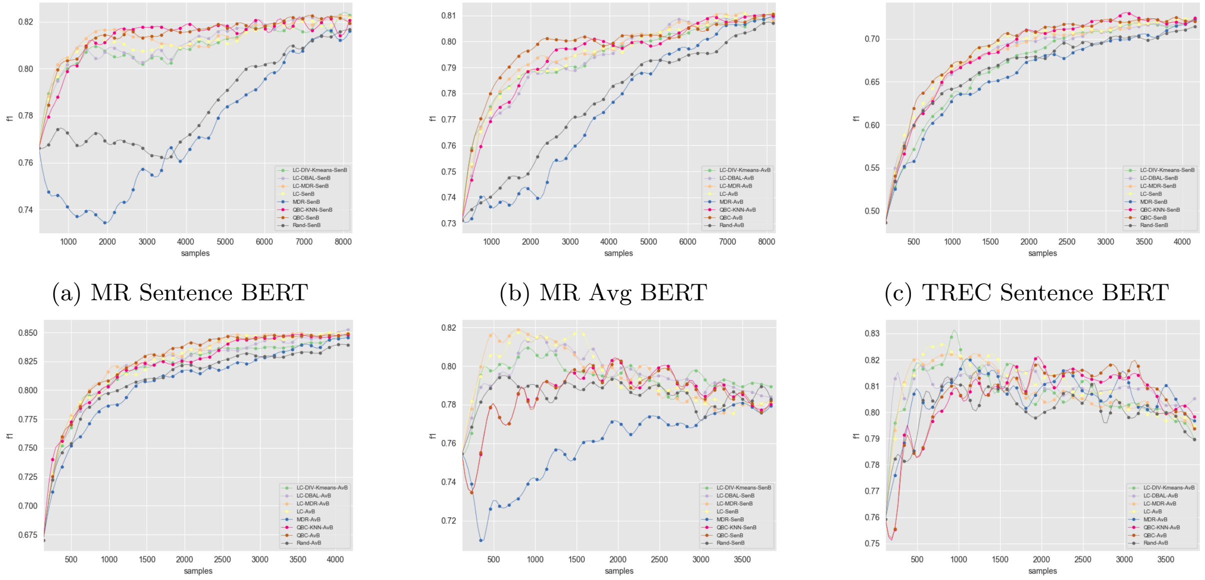
We used 5 fold cross validation and split the data into train and test sets. At every fold our experiments considered all possible pairs combinations between sentences' representations and query strategies. We used the following query strategies : Least Confidence (LC) , Least Confidence Diverse by K means (LC-DIV-Kmeans) , Least Confidence Diverse Batch Active learning (LC- DBAL), Least Confidence Max Diverse and Representative (LC-MDR), Max Diverse and Representative (MDR),Query-By-Committee(QBC), Query-By-Committee k-nearest neighbours (QBC-KNN) and random sampling (Rand). We simulated pool-based active learning with fully labeled datasets. For each iteration of training the number of smaples to query was determined to be 3% of the total training samples. The AL process started with random i.i.d samples, then the AL loop began and the samples were chosen by the relevant query strategy. The AL loop ended when all of the training data was sampled.

5.3 Evaluation

Classification model performance evaluated with F1 score at each training iteration step

[1] https://github.com/UKPLab/sentence-transformers

[2] https://github.com/hanxiao/bert-as-service



(a) MR Sentence BERT

(b) MR Avg BERT

(c) TREC Sentence BERT

(d) TREC Avg BERT

(e) TOXIC Sentence BERT

(f) TOXIC Avg BERT

Figure 1: F1 learning curve for MR (a,b), TREC (c,d), TOXIC (e,f) datasets and for the different representations: Sentence BERT(a,c,e), Avg BERT (b,d,f).

for each query strategy. The baseline query strategy is the random sampling where in each step n random samples picked from the unlabeled pool. For each dataset we generate progress record of F1 changes by training iteration step. The area under that learning curve (ALC) is calculated according to (Guyon et al., 2011).

$$ALCscore = \frac{ALC_{strategy} - ALC_{random}}{ALC_{max} - ALC_{random}} \tag{11}$$

Where ALC_{max} in our experiment is the area under the maximum F1 curve ,with F1 of 1 . We scaled each ALC to the ALC of the random query strategy.

6 Results

6.1 Active learning increase labelling efficiency

Although published in many papers (Tang et al., 2002; Garg and Sundararajan, 2009; Laws et al., 2011; Li et al., 2012a) ,it is worth to mentioning that using active learning with several of the methods we have suggested reached maximal F1 with less annotated samples. It is highly visible in table 2 and figure 1 in the MR dataset where except MDR ,all other query strategies increased the area under the F1 learning curve. In TREC we also see increase in the ALC for some of the methods , such as LC and QBC but the effect is smaller. Toxic dataset results are less conclusive due to large variance between the folds where in some we see ALC improvement and in some not at all.When we analyze Toxic performance we see that the maximal F1 was reached in early steps of the AL process, and indeed when re-measuring the ALC at the 10th iteration, i.e after 1200 samples, we can see ALC improvement especially in LC-MDR with 0.09 ± 0.04 and 0.07 ± 0.03 in SentenceBert and AVGBert respectively.Therefore we can see that Active learning can indeed reduce the amount of annotated labels needed for higher learning performance but the stability of the methods is not guaranteed.

6.2 Comparison between AL methods

In our experiments we combined both model-naive and model-dependent methods together. Here we will divide them into two groups based on the usage of the representation component: representation-based (Rep) which includes LC-DIV-Kmeans, QBC-KNN ,MDR, LC-MDR, LC-DBAL. In the No-representation group (NoRep) we included LC and QBC methods.Table 2 shows that the differences in the ALC between the NoRep group and the corresponding methods with added representation in the Rep group are minor and can

Representation	Sample Method	MR	Toxic	TREC
SenB	LC-DIV-Kmeans	0.113±0.02	0.035±0.06	0.004 ±0.02
	LC	0.122 ±0.02	0.04 ±0.06	0.054 ±0.02
	QBC	**0.136 ± 0.01**	-0.004 ±0.06	0.065 ±0.02
	QBC-KNN	**0.132 ± 0.02**	-0.004 ±0.06	0.049 ±0.02
	MDR	-0.056 ±0.03	-0.127 ±0.08	-0.029 ±0.02
	LC-MDR	0.13 ±0.02	**0.041 ± 0.04**	0.041 ±0.03
	LC-DBAL	0.121 ±0.02	**0.043 ± 0.05**	0.038 ±0.03
	Rand	0	0	0
AvB	LC-DIV-Kmeans	0.077 ±0.02	0.03 ±0.05	0.044 ±0.03
	LC	0.081± 0.02	0.038± 0.07	0.068± 0.02
	QBC	0.094± 0.02	0.02± 0.06	**0.08± 0.03**
	QBC-KNN	0.079± 0.02	0.017± 0.06	0.059± 0.01
	MDR	-0.012± 0.02	0.021± 0.06	-0.019± 0.02
	LC-MDR	0.086± 0.02	0.031± 0.05	**0.082± 0.02**
	LC-DBAL	0.076± 0.02	0.04± 0.04	0.056± 0.01
	Rand	0	0	0

Table 2: Summary of the ALC score for each query strategy in every dataset

not be interpreted as significant. These results are surprising as they are opposed to previous works (Zhu et al., 2008; Huang et al., 2010; Zhdanov, 2019). one possible explanation is that the model does not fully utilise the text representation. We analyzed the chosen samples at the first iteration of the AL on MR dataset. figure 2 visualize the chosen samples' representation with T-SNE, the colors indicate the query strategy (MDR, LC, LC-MDR and random).The samples chosen by the MDR are much more sparse and far away from the samples chosen by LC , which are grouped around the center of the space. Since the LC strategy represents the model's uncertainty it might indicate that the MDR samples will contribute less to the model's certainty. In addition, samples chosen by the MDR strategy are even more sparse than samples chosen by the random baseline. It might be the reason for the poor performance of the MDR strategy compare to the random baseline. The samples chosen by LC compared to LC-MDR are very close to each other, it can explain the minor differences at the ALC scores between them.

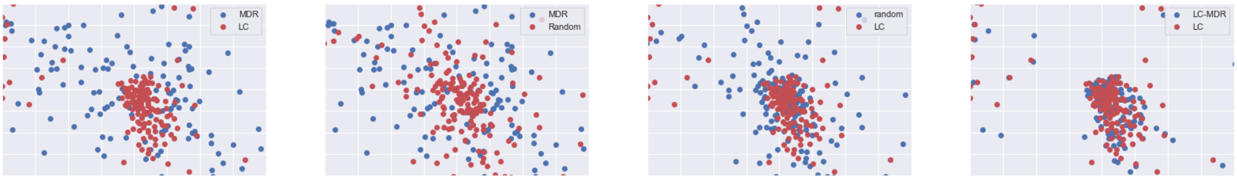
6.3 The influence of the representation on the AL

To assess whether there are differences in the impact of query methods , representations and the overall ALC improvement , we compared the ALC changes for each dataset.In MR and Toxic both representations reached very similar F1 scores, but we can see that the ALC improvement was higher in Sentence Bert. In TREC the F1 differences were much higher, 0.84 in AVG Bert and 0.71 in Sentence Bert and indeed the highest ALC improvement was in AVG Bert. We suggest that the ALC improvement is not related only to how well the model can learn the classification task but to how well the embedding space can represent the differences between each sentence. We can see in figure 3 for an example that TREC Sentence Bert representation does not show separated classes in comparison to AVG Bert representation , where samples of the same class are clustered together. In MR we see similar F1 results but the ALC improvement are much larger in Sentence Bert and respectively the t-SNE visualisation map depict for Sentence Bert a much better class separation.

7 Discussion

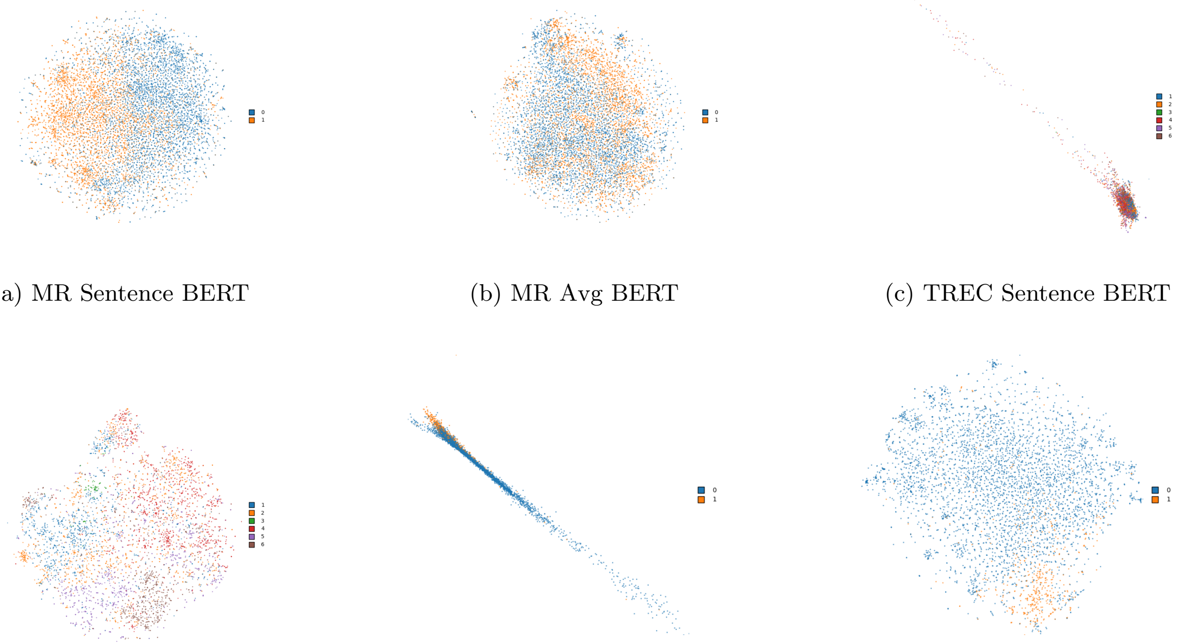
We have investigated the behavior of several active learning methods that combine representation and uncertainty sampling with two sentence embedding methods. The primary research question was dealt with the relation between the text representation and the effectiveness of the active learning procedure. We have shown that the potential in the effectiveness is not fully determined by the type of the embedding method and not by the model performance at the end point. We suggest that the magnitude of the effectiveness in using AL is derived by how well each class can be separated in the embedding space. When We examine the t-SNE plots (figure3) of MR sentence Bert we see a good separation between the samples of each class in the perimeter and overlapping of the two clusters in the center. We suggest that the center is the ”uncertainty zone” between those samples , and indeed when we plotted the t-SNE of samples queried by the least confidence method , we see the samples accumulated in the center in comparison to the samples queried randomly. The weak relation between the success of the AL and representation based query strategies might reveal a model weakness in the representation utilization.

Future studies may focus on the performance of different AL methods using models such as neural networks that are capable to utilise better the representation .Better understanding of the sub groups in the uncertainty zones might provide us the ability to define better methods that can query the optimal set of samples in each iteration.



(a) LC vs MDR (b) Random vs MDR (c) LC vs random (d) LC vs LC-MDR

Figure 2: T-SNE visualisations of the chosen samples at the first AL iteration from MR dataset. The colors indicate the query strategy, (a) samples chosen by LC strategy compare to MDR strategy, (b) samples chosen by MDR strategy compare to random baseline, (c) samples chosen by LC strategy compare to random baseline and (d) samples chosen by LC strategy compare to LC- LC-MDR stratey .



(a) MR Sentence BERT (b) MR Avg BERT (c) TREC Sentence BERT

(d) TREC Avg BERT (e) TOXIC Sentence BERT (f) TOXIC Avg BERT

Figure 3: T-SNE visualisations of MR (figure 2(a),2(b)), TREC (figure 2(c),2(d) and TOXIC (figure 2(e),2(f)) datasets. The colors indicate the class label.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Rosa Figueroa, Qing Zeng-Treitler, Long Ngo, Sergey Goryachev, and Eduardo Wiechmann. 2012. Active learning for clinical text classification: Is it better than random sampling? *Journal of the American Medical Informatics Association : JAMIA*, 19:809–16.

Priyanka Garg and S. Sundararajan. 2009. Active learning in partially supervised classification. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 1783–1786, New York, NY, USA. Association for Computing Machinery.

Isabelle Guyon, Gavin C Cawley, Gideon Dror, and Vincent Lemaire. 2011. Results of the active learning challenge. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, pages 19–45.

Jeremy Howard and Sebastian Ruder. 2018. Fine-tuned language models for text classification. *ArXiv*, abs/1801.06146.

Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. 2010. Active learning by querying informative and representative examples. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, page 892–900, Red Hook, NY, USA. Curran Associates Inc.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML'98, page 137–142, Berlin, Heidelberg. Springer-Verlag.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura E. Barnes, and Donald E. Brown. 2019. Text classification algorithms: A survey. *Information*, 10:150.

Florian Laws, Christian Scheible, and Hinrich Schütze. 2011. Active learning with Amazon mechanical turk. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1546–1556, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Lianghao Li, Xiaoming Jin, Sinno Jialin Pan, and Jian-Tao Sun. 2012a. Multi-domain active learning for text classification. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 1086–1094, New York, NY, USA. Association for Computing Machinery.

Shoushan Li, Shengfeng Ju, Guodong Zhou, and Xiaojun Li. 2012b. Active learning for imbalanced sentiment classification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 139–148, Jeju Island, Korea. Association for Computational Linguistics.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Jinghui Lu, Maeve Henchion, and Brian Mac Namee. 2019. Investigating the effectiveness of representations based on word-embeddings in active learning for labelling text datasets. *ArXiv*, abs/1910.03505.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, page 589–es, USA. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? *ArXiv*, abs/1905.05583.

Min Tang, Xiaoqiang Luo, and Salim Roukos. 2002. Active learning for statistical natural language parsing. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 120–127, USA. Association for Computational Linguistics.

Fedor Zhdanov. 2019. Diverse mini-batch active learning. *CoRR*, abs/1901.05954.

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, COLING '08, page 1137–1144, USA. Association for Computational Linguistics.