## Proposal Notes

*Traditional Bayesian AL methods use uncertainty sampling (i.e. informativeness is measured by predictive uncertainty) and typically require probabilistic machine learning models to acquire good uncertainty estimates for the candidate data points. However, current work uses deep learning models that provide large performance gains but not well-calibrated confidence scores (Guo et al., 2017), i.e. predictive softmax probabilities are erroneously interpreted as model confidence (Gal and Ghahramani, 2016). Several approaches have been proposed to calibrate the output probability distribution of deep neural networks, such as temperature scaling (Guo et al., 2017), Monte Carlo dropout (Gal and Ghahramani, 2016) and model ensembles (Lakshminarayanan et al., 2017). Using uncertainty sampling with the vanilla output probabilities for AL may lead to incorrect conclusions, i.e. poor results may be attributed to the acquisition method, while the problem may be in fact the lack of calibration. Still, only a few deep Bayesian AL approaches apply a calibration method to the posterior probabilities (Gal et al., 2017; Shen et al., 2017; Siddhant and Lipton, 2018; Lowell and Lipton, 2019; Ein-Dor et al., 2020)*

This work focused on various binary classification tasks. A natural future direction is to conduct

a similar empirical investigation of AL over BERT in the context of multi-class classification and regression tasks. The development of novel AL methods, that are tailored for pre-trained models such as BERT, seems like an important direction for future work.

It would also be interesting to investigate the realm of larger annotation budgets, and more recent BERT variants.

To address this gap in the literature this paper addresses four research questions:

To answer these research questions, this paper describes a comprehensive evaluation experiment that ex

investigate the effectiveness of various BERT-like models and identify the appropriate representation method for

pre-trained language models.

We perform evaluations on three language understanding tasks:

Mentioned any work doing AL with BERT/transformers.

- What worked, what didn't?  
  
- Based on that, what is the starting point for your thesis work? What are the main challenges for AL with transformers? How do you want to address them?  
  
In my opinion, the two main challenges for your work are:  
- class imbalance when focusing on a multi-class classification of transactions, where a single majority category drives the transaction.

-the high time requirement for training transformers

- the fact that softmax is not a good predictor for model confidence

Several approaches have been proposed to calibrate the output probability distribution of deep neural networks, such as temperature scaling (Guo et al., 2017), Monte Carlo dropout (Gal and Ghahramani, 2016) and model ensembles (Lakshminarayanan et al., 2017). Using uncertainty sampling with the vanilla output probabilities for AL may lead to incorrect conclusions, i.e. poor results may be attributed to the acquisition method, while the problem may be in fact the lack of calibration.

One step towards understanding whether these models can be trusted is by analyzing whether they are calibrated (Raftery et al., 2005; Jiang et al., 2012; Kendall and Gal, 2017): how aligned their posterior probabilities are with empirical likelihoods

We improve data acquisition by providing well-calibrated uncertainty estimates by using Monte Carlo dropout (Gal and Ghahramani, 2016) instead of using the softmax output as confidence scores <https://arxiv.org/pdf/2104.08320.pdf>

How to address them?:

. Nonetheless, our results show that representations induced by robust pre-training (e.g., using a larger corpus, more training steps, dynamic masking) (Liu et al., 2019) lead to more calibrated posteriors. <https://aclanthology.org/2020.emnlp-main.21.pdf>

(Deep AL presents some specific challenges.

Since DNNs are computationally heavy, training

a new model whenever a single training sample is

added is highly impractical. This requires a shift

to batch mode active learning, where a batch of examples

is queried at every iteration. Moreover, the

tendency of the softmax layer to over-confidence

has led to the development of various uncertainty based

strategies tailored to the special properties of

DNNs (Gal and Ghahramani, 2016).

Most of the works in deep active learning)

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a

bayesian approximation: Representing model uncertainty

in deep learning. In international conference

on machine learning, pages 1050–1059.

https://arxiv.org/pdf/2011.06225.pdf  
  
It would be good if you could discuss those challenges (and maybe you'll find even more) in the proposal and say how you want to approach them. Of course you won't be able to solve everything in one thesis, but it would be good to be a bit more specific about which problems you want to investigate and how.

## Uncertainty in NLP

* However, current work uses deep learning models that provide large performance gains but not well-calibrated confidence scores (Guo et al., 2017), i.e. predictive softmax probabilities are erroneously interpreted as model confidence (Gal and Ghahramani, 2016). Several approaches have been proposed to calibrate the output probability distribution of deep neural networks, such as temperature scaling (Guo et al., 2017), Monte Carlo dropout (Gal and Ghahramani, 2016) and model ensembles (Lakshminarayanan et al., 2017)
* [We compare two different uncertainty measures for our method. The first is entropy: H(p(y|x)) = −Ep(y|x) [log](http://bayesiandeeplearning.org/2018/papers/6.pdf) p(y|x)] and the second is the Bayesian active learning by disagreement (BALD) uncertainty measure [4]:

For datasets: the options you listed sound good to me.

## You might want to check that the datasets you choose:

- are large enough so that you have enough instances to use as pool of "unlabelled" data for AL

- have information on inter-annotator agreement (IAA) between the human annotators when creating the dataset: the more reliable the annotations, the better AL will work. On the other hand, it might also be interesting to consider IAA as a variable that you control for by also picking some datasets with a lower IAA and then evaluate the impact on AL.

- the length of the instances (text length) is also an important variable. You will probably want smaller instances that will fit the BERT sequence length of 512 wordpiece tokens (sentences and short paragraphs but not whole documents)

- the number of labels (in general: the more labels, the more difficult the task; but this also depends on the task and how ambiguous the labels are, which you can check by looking at the IAA for the annotations)

- the distribution of the labels (data imbalance is a challenge for AL but is very common in real-world datasets)

- the language (do you only want to work with English or do you want to test on other languages, too?)

In short, there are many variables that you might want to consider. You need to think about it and make a decision which ones you want to investigate (and select datasets that differ with respect to these variables) and which ones you don't want to investigate (then you might want to control for these variables by chosing datasets that are similar with reard to this variable). These are already important decisions for your thesis work and should be motivated by the research questions you want to investigate in your thesis.

We used eight, large, representative datasets widely used for text classification**:**

**http://nlpprogress.com/english/text\_classification.html**

1. **AGNews (AGN),**
2. [**DBPedia**](https://datasetsearch.research.google.com/search?query=multi%20class%20nlp&docid=L2cvMTFqY2p4NGZrNw%3D%3D) **(DBP)**,
3. **TREC**
4. EURLEX57K (19.6k documents, 4k [EUROVOC](https://arxiv.org/pdf/1906.02192v1.pdf) labels) - Large-Scale Multi-Label Text [Classification on EU Legislation](https://github.com/iliaschalkidis/lmtc-eurlex57k) <https://aclanthology.org/W19-2209.pdf>

https://www.uco.es/kdis/mllresources/

1. [Germeval 2019](https://projects.fzai.h-da.de/iggsa/projekt/), which consists of [German tweets](https://towardsdatascience.com/bert-text-classification-in-a-different-language-6af54930f9cb)
2. ML-NET https://github.com/jingcheng-du/ML\_Net-1

**Other datasets: https://www.uco.es/kdis/mllresources/#EukaryoteDesc**

Amazon Review Polarity (AMZP), Amazon Review Full (AMZF), Yelp Review Polarity (YRP), Yelp Review Full (YRF), Yahoo Answers (YHA) and Sogou News (SGN)  
We finally use the large-scale AGNEWS and DBPEDIA datasets from Zhang et al. (2015) for topic classification. The first task is question classification using the sixclass version of the small TREC-6 dataset of opendomain, fact-based questions divided into broad semantic categories (Voorhees and Tice, 2000).

https://gscl.org/en/arbeitskreise/cpss/cpss-2021/workshop-programme#poster5

After reading Schröder et al. 2021, I have some more comments.

In my opinion, they ignore relevant issues for AL:

- They only run simulation studies with no real human annotators in the loop and ignore long training times needed for the larger models (you can't hire an annotator, let them annotate 25 instances and then wait for 10 minutes or until the next model is trained. You will also run simulations, but we should take care that our models don't need training times that would make them infeasible to use in practice.

- The improvements over the random baseline are very small, and no significance testing has been done. I don't think that the paper has managed to show the advantages of AL.

- The paper only uses balanced datasets (with one exception: TREC-6, but they give no information on the class distribution), which is highly unrealistic and will give overly optimistic results.

- The paper only uses "easy" tasks where the random baseline already achieves accuracies >90%. AL is much harder for tasks that are more ambiguous, where model scores are much lower (and also the scores for human annotators).

Based on those observations, I think it would be interesting for the thesis to focus on

- Methods that don't need a large amount of training/selection time (to make sure that the method can be applied in a real-world scenario)

- Instead of trying many different sampling methods, I would rather focus on investigating the impact of language, number of classes, data imbalance and difficulty of the task (e.g., measured in terms of classification score for the random baseline) on AL performance. I think this would make a more interesting (and realistic) contribution than testing a multitude of complex sampling methods on easy datasets, as done in the paper ;) So I would advice you to check for each sampling method that you want to try if it's time requirements are realistic/small enough to be used in real-world applications.

To do it, we proposed a new method, sampling by uncertainty and density (SUD), in which entropy-based uncertainty measure and KNN-density measure are considered simultaneously. In SUD scheme, a new uncertainty measure, called density\*entropy measure 4 , is defined as: ×= xHxDSxDSH )()()(

In this section we utilize an approach, sampling by clustering (SBC), to selecting the most representative examples to form initial training data set. In the SBC scheme, the whole unlabeled corpus has been first clustered into a predefined number of clusters (i.e. the predefined size of the initial training set). The example closest to the centroid of each cluster will be selected to augment initial training set, which is viewed as the most representative case. ( C08-1143.pdf)

~~- Of course you can try all pretrained transformer models that you'd like to, I would also advice to include a smaller, leight-weight model such as DistilBert or DistilRoberta, to reduce runtime.~~

~~In addition, I don't think that softmax is an issue (it is for word embedding training where the number of classes equals the vocabulary size (which is some value between 10,000 and 50,000), and therefore Mikolov et al (2013) used hierarchical softmax for their skipgram model. I don't think that for class sizes <20, this is a problem. If you know better, let me know :)~~

I would suggest that we have a meeting to discuss my suggestions once I'm back from holidays (beginning of Sep) and then you finalize the proposal and register. Of course, you can already start working on the thesis as the proposal is nearly done.

# Literature Review - Thesis

## Important links & Papers

|  |  |  |
| --- | --- | --- |
| Title | Purpose | Location |
| Thesis-Towards Practical Active Learning for Classification |  | Literature\_review |
| Investigating the Effectiveness of Representations Based on Pretrained Transformer-based Language Models in Active Learning for Labelling Text Datasets |  | Literature\_review |
| A SURVEY OF ACTIVE LEARNING FOR TEXT CLASSIFICATION  USING DEEP NEURAL NETWORKS |  | Literature\_review |
| Active Learning for BERT An Empirical Study |  | Literature\_review |
| Bayesian Active Learning with Pretrained Language Models |  |  |
| Uncertainty-based Query Strategies for Active Learning with Transformers |  | Literature\_review |
| From Theories to Queries: Active Learning in Practice |  | Literature\_review |
| MII: A Novel Text Classification Model Combining Deep Active Learning with BERT |  |  |
| **Active Learning Literature Survey** |  | Literature\_review |
| A Review of Uncertainty Quantification in Deep  Learning: Techniques, Applications and  Challenges |  |  |
| Calibration of Pre-trained Transformers |  |  |
| The Application of Active Query K-Means in Text Classification |  |  |
| Active Learning with Sampling by Uncertainty and Density for Word Sense Disambiguation and Text Classification | SUD & SBC |  |
| Initial Training Data Selection for Active Learning | Fuzzy clustering algorithm |  |
| Active Learning Strategies for Semi-Supervised DBSCAN (book,  pp 179-190) |  |  |
| Text Classification Algorithms: A Survey |  |  |

## Section: Introduction

* ~~Intro & Definition~~

## Intro to to AL

* **~~AL & 3 main scenarios~~**

*Over the course of one AL run, an agent annotates its dataset exhausting its labeling budget. Thus, given a new task, an active learner has no opportunity to compare models and acquisition functions.*

*Strategically selecting points to annotate over alternating rounds of labeling and learning, an active learner is hoped to outperform budget-matched i.i.d. labeling. Typical acquisition functions select examples for which the current predictor is most uncertain. (cite\* Deep Bayesian Active Learning for Natural Language Processing: Results of a Large-Scale Empirical Study)

In active learning, the learner has the ability to query the labels of instances. There are three are main scenarios which are illustrated in Figure 1 and include:

* Membership Query Synthesis, where the learner generates its new artificial instances from an underlying natural distribution. \cite{a-qcl-88} This scenario is reasonable for many finite problem domains (Angluin, 2001). Membership query synthesis for natural language processing tasks may create streams of text that produces gibberish and labeling such arbitrary instances can be awkward when the oracle is a human annotator. \cite{settlestr09}
* Stream-Based Selective Sampling, a learner receives distribution information from the environment

and queries an oracle on parts of the domain it considers "useful"\cite{cohn:94a}, i.e the unlabeled instances are queried one at a time and active learning algorithms have to decide whether or not to ask a human expert to label it.

The approach not only reduces annotation effort, but also limits the size of the database used in nearest-neighbor learning, which in turn expedites the classification algorithm

*If the input distribution is uniform, selective sampling may well behave like membership query learning. However, if the distribution is non uniform and (more importantly) unknown, we are guaranteed that queries will still be sensible, since they come from a real underlying distribution.*

* Pool-Based Sampling, where the instances are selected from the entire data pool, which is assumed to which is usually assumed to be closed (i.e., static or non-changing), based on an informativeness measure used to evaluate all the instances in the pool. [David D. Lewis 1994]

Stream-based active learning algorithms can perform worse than pool-based methods. *\cite{* *pmlr-v22-ganti12}* More data points are queried for human annotation in the stream-based setting than that in the pool-based methods. One reason is that in the stream-based setting, AL algorithms cannot go through all the unlabeled data to select the most useful samples. It is likely that before the most informative samples are presented in a stream, the annotation budget is already finished. Most active learning algorithms focus on the pool-based scenario. In this thesis, I will mainly concentrate on exploring new strategies for the pool-based setting.

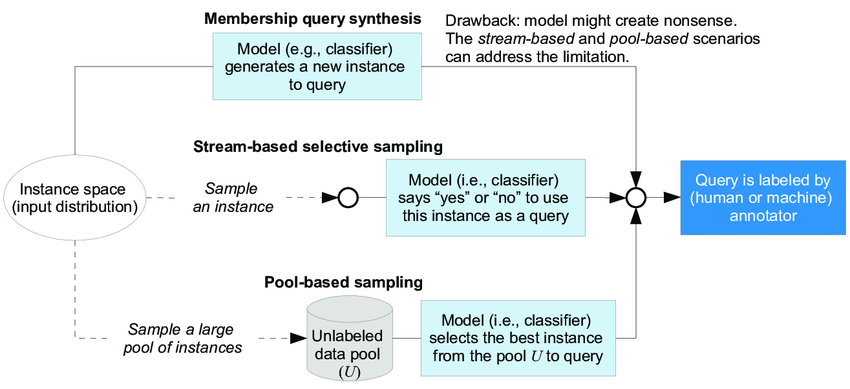
The advantage of stream-based strategies lies on its computational efficiency, as there is no need to go through the data pool to query the best sample. This approach reduces annotation effort and limits the size of the database used in nearest-neighbor learning which expedites the classification algorithm making it more efficient. \cite{settlestr09}The price of high efficiency is a weaker performance.

*(Ganti and Gray found that stream based active learning algorithms are likely to perform worse than pool-based methods, for instance, more data points are queried for human annotation in the stream-based setting than that in the pool-based setting. One reason is that in the stream-based setting, active learning algorithms cannot go through all the unlabeled data to select the most useful samples. It is likely that before the most informative samples are presented in a stream, the annotation budget is already finished. Most active learning algorithms focus on the pool-based scenario [cite: X. Li and Y. Guo, Adaptive active learning for image classification, in Proceedings of*

*the IEEE Conference on Computer Vision and Pattern Recognition (2013) pp. 859–*

*866., cite: J. Cheng and K. Wang, Active learning for image retrieval with co-svm, Pattern*

*recognition 40, 330 (2007)., cite: S. Tong and D. Koller, Support vector machine active learning with applications to text classification, The Journal of Machine Learning Research 2, 45 (2002).,thesis references page 3. 40–46]. In this thesis, I will mainly concentrate on exploring new strategies for the pool-based setting.)*



* **~~Sampling Strategies~~**
* **Dig into uncertainty, i.e. extend the work done with the different uncertainty methods and show the gap in the literature**.

The different methods of active learning start by measuring the informativeness of unlabeled instances and select(sample) new instances that support the learning process. There are various selective sampling frameworks for active learning. Amongst the most popular ones are:

* + Query-by-committee (QBC) selects a committee of classifiers by randomly sampling hypotheses from the version space. [21]  
    QBC constructs a committee consisting of a number of different classifiers/models (always more than two). The next unlabeled example is selected by the principle of maximal disagreement among these classifiers, i.e., each committee member can vote on the labeling of unlabeled instances and the instance which causes greatest disagreement within the committee will be selected for annotation.
  + Expected Model Change- identifies the instance that would impart the greatest change to the current model if we knew its label. \cite{Cai\_wenbin}

*The intuition behind this framework is that it prefers instances that are likely*

*to most influence the model (i.e., have greatest impact on its parameters), regardless*

*of the resulting query label. This approach has been shown to work well in*

*empirical studies, but can be computationally expensive if both the feature space*

*and set of labelings are very large. (cite settles )  
The differences among these approaches lies in the criteria to measure*

*the model change. Settles et al. [77] proposed to measure the expected gradient*

*length of the objective function. Freytag et al. [79] estimated the change of model outputs*

*instead of model parameters.*

*Cai et al. [45] proposed to use the gradient of the loss function to approximate the*

*model change and derived algorithms for both SVM and logistic regression classifier.*

* + Uncertainty Sampling- in this framework, an active learner queries the instances about which it is least certain how to label. [15]  
    *(This approach is often straightforward for probabilistic learning models.)*

*The motivation behind uncertainty sampling is to find some unlabeled examples near decision boundaries and use them to clarify the position of decision boundaries.*

Predictive uncertainty in deep neural generally consists of two parts: data uncertainty (aleatoric) \cite{2cfc44fb6aa842f5ac03e51409667171}and model uncertainty(epistemic). \cite{DBLP:journals/corr/KendallG17} *Defining the right acquisition function, i.e., the condition on which sample is most informative for the model, is the main challenge of AL-based methods. For AL*

*Uncertainty measures:  
Entropy:* A more general uncertainty sampling strategy uses entropy (Shannon, 1948)

as an uncertainty measure. Entropy is an information-theoretic

measure that represents the amount of information needed to “encode” a distribution. As such, it is often thought of as a measure of uncertainty or impurity in machine learning.

For binary classification, entropy-based uncertainty sampling is identical to choosing the instance with posterior closest to 0.5. However, the entropy-based approach can be generalized easily to probabilistic multi-label classifiers and probabilistic models for more complex structured instances, such as sequences (Settles and Craven, 2008)

GAP: *Active learning (AL, also sometimes called optimal experimental design) plays a key role in dealing with lack of labeled data to label new data with the desirable outputs. As most researchers in the domain of machine and deep learning are aware that having labeled data is very difficult, costly and time consuming. In few very extremely sensitive areas such as healthcare and self-driving cars, this importance becomes more apparent. We have reviewed several good studies in this area, including: Gordonet al. [561], Lee et al. [311], Nguyen et al. [562], Hu et al. [563], Qu et al. [564] and some few more. Our reviews; however, reveal that although uncertainty in this area is quite important, but very few studies have been done in this subject. For example, Sinha et al. [565] proposed a new Al in an adversarial manner which is called VAAL (Variational Adversarial AL). The proposed model trained a latent space by using a VAE and an adversarial network trained to discriminate between labeled and unlabeled data (see Fig. 24). They showed the dramatic impact of this type of method on UQ. For this reason, we believe that researchers can fill this gap with further studies in order to improve the data labeling quality by having far more certainty than previous studies. { a review of uncertainty quantification in*

## Section: Advances in Text Classification

Natural language processing models have achieved state-of-the-art results using transformer based machine learning like BERT (Bidirectional Encoder Representations from Transformers). The first empirical study on the combination of AL strategies with BERT was conducted by Ein-Dor et al. [8] The focus of their work were various binary classification tasks, and to understand the impact of AL in the performance when using BERT based models they explored several AL strategies:

1. Random was the baseline - randomly sampling from unlabeled data
2. Uncertainty-based sampling (Least Confidence, Monte Carlo Dropout (Gal et al., 2017))
3. Uncertainty-sampling using ensemble methods (Perceptron Ensemble)
4. Expected Model Change (Expected Gradient Length (EGL) - choosing the examples with the largest expected gradient norm, with the expectation over the posterior distribution of labels for the example according to the trained model. [13])
5. Diversity Sampling (Discriminative Active Learning (DAL)- selecting instances that make the labeled set of instances indistinguishable from the unlabeled pool. [10],
6. Core-Set - choosing the examples that best cover the dataset in the learned representation space using farthest first traversal algorithm. [18])

## They performed a comparative analysis between the different strategies by measuring two batch properties that are known to impact the AL effectiveness, diversity and representativeness. AL strategies showed to further boost the performance of BERT, however an inconsistency in the performance of AL was observed, leaving an open question: Which selection strategies can improve the classification even further?

## Transformer-based language models like BERT have been pre-trained on large text corpora, and they can be fine-tuned to a specific task using less training data than when trained from scratch. These models have proven to be highly performant but they have a high number of model parameters which makes them computationally very expensive for query strategies that are targeted at text classification.

## Section: My Contribution

assume a pool-based batch-mode scenario because in a text classification setting the dataset is usually a closed set, and the batch-wise operation reduces the number of retraining operations, which cause waiting periods for the user.

## Section: Theoretical Framework

**Rocco tips:**

1. Synthetically generate more data and train on it
2. Get subset of data using regex, and then use a xgboost or non-neural network modelextract relevant portions of the contract using regex
3. vectorize the text using a simple model like tf-idf
4. train a non deep learning classifier like XGBoost
5. Try some deep learning models, this time on word embeddings

<https://dash.plotly.com/>

https://github.com/plotly/dash-sample-apps/