Understanding the Effectiveness of Active Learning in NLP in a Transfer Learning setting

Pawan Kumar Singh and Abinaya Mahendiran

I. Motivation

With the advent of EMLO and ULMfit, began the transfer learning wave in NLP, and this in turn has made model building easier for new datasets. However, when it comes to the representation of spoken languages in the world, around 50% of datasets used in NLP are in English, with next in line being Chinese (6%) and German (6%) [1]. Languages spoken by our team, - Hindi and Tamil, have around 2% and 1% of the dataset. If one digs deeper, most of the publicly available Hindi dataset is part of a multilingual dataset and there are hardly any known and famous benchmarks on them. If models are built on datasets other than English, it would benefit a very large section of the society.

II. Objective

The goal is to understand how models can be built for languages where the publicly available datasets are less. With this goal in mind, we set out to discover how we can rapidly generate high-quality datasets for English using active learning (AL) in a transfer learning setting. To do so, we chose a very well studied AG's news corpus dataset [2] (English) on which multiple benchmarks are available. This dataset is mostly used to build model to classify text into four classes. Our goal is to determine how much effort would be required to create a high-quality dataset without having to manually tag a large number of data points to begin with, using active learning in a transfer learning setting. Specifically, we try to answer the following questions:

- What is initial data size, which can be used to leverage active learning while harnessing the benefit of transfer learning ?
- At what data size, gains from active learning become hard to realize?
- Which sampling method works best for NLP?

While trying to find the answers to the above questions, we showcase the challenges faced, what is done to overcome them and how the best practices learned in the class and from [7], are applied to the entire process. We end our discussion with why annotation is hard, especially in NLP and what all features an ideal annotation tool should have from our experience working on this project.

III. Data

AG's news corpus dataset [2] is used, which contains 496,835 categorized news articles from more than 2000 news sources. We use a part of the dataset available on kaggle [3], which has been constructed using the title and description of four largest classes from the original corpus (which has around 1 million data points). The four classes are:

1) World News
2) Sports News
3) Business News
4) Science and Technology News

Each classes have 30,000 training and 1,900 testing sample, summing to 120,000 training and 7,600 testing samples.

IV. Model

For the task of classifying the news articles, multi-class classification, into one of the four classes, we used a pre-trained RoBERTa [6] (roberta-base) model trained on a large English corpus. It is done using Simple Transformers [4] library, which is a wrapper over the famous Hugging Face Transformers [5] library. The "roberta-base" is fine-tuned for the classification task using some default parameters (model arguments), except for the training epochs, which was chosen on the basis of the performance on baseline data. Since, we are leveraging transfer learning, fine-tuning for high numbers of epochs leads to high variance. So, the number of epochs is set to five.

V. Experiments

The total training samples (120,000) are split into training, validation and annotation data for experimenting with active learning. The details of the split for each of the experiments are provided in Fig. 1. The size of annotation data and validation data is fixed, in order to avoid the effect of different data being sampled across different experiments. Size of the training data varied from 1,000 to 6,000 to 30,000 samples across different experiments. All the training was done on Google Colaboratory. Total testing samples (7600) is used to assess the performance of the retrained model using the annotated samples.

To demonstrate active learning, the following sampling methods, based on the uncertainty of the model predictions, have been adopted in the experiments to select data for annotation:

- Random Sampling
- Least Confidence Sampling
- Entropy Based Sampling

We intended to perform a total of 13 experiments that included a baseline and experiments where all the three sampling methods are used across three different data regimes, Fig 1. The baseline experiment and the experiments related to low and high data regime are completed. Medium data regime

experiments are not performed because of the time crunch and annotation complexity. In each of the experiments, the first step is to train the "roberta-base" model on the train and validation split and then re-train on the annotated samples using the specified method. 1000 and 200 samples were annotated for high and low data regime respectively.

In an initial experiment, the train split is fixed as 25% of the training samples (120,000) and the "roberta-base" model is fine-tuned. The idea is taken from the project suggestion documentation from the course website. But, when the model is re-trained using the annotated samples (1000) for each of the sampling methods separately, we did not observed any improvement in the test accuracy. Because of poor generalization, we decided to carry out experiments with different data regimes, where the train split used for initial training is different across the experiments, Fig 1. This resulted in better generalization, improved test accuracy, but led to other problems, as discussed in the results section. The idea of using different data regimes (low, medium and high) stemmed from the real world experience, where the availability of tagged data is minimal. Because of the time constraints, experiments related to medium data regime are not carried out.

A custom command line tool was built for data annotation. This tool has ability:

- to undo any annotation
- assign a **Not Sure** class in case of confusion, rather than assigning them to **World News**
- see randomly sampled examples for any class when in confusion while annotating

Though this tool served us well, but its not very easy to use and lacked some of the features which we would like to have, discussed in **Annotation is hard** section. Hence, we started building a dash based tool, which is work under progress at the time of writing this report.

VI. RESULTS

In this section, we present the results of each of the experiments, while comparing the performance of different sampling methods. For each data regime and the sampling method combination, there are two baselines to compare:

- First, **overall baseline** is the model that is trained on the available training samples, 120,000 of them with 80-20 train and valid split and tested on the testing samples. This remains the same for all the data regimes.
- Second, **baseline** model is the one that is trained on data regime specific training samples and tested on the original testing samples. Each data regime has its own baseline.

Test set accuracy is the metric of choice to compare all sampling methods among themselves and with the baseline.Table I provides the accuracy of the model re-trained using different sampling method for the high data regime, whereas table II provides the details for low data regime.

From Table I, it becomes clear that if we have sufficiently large number of data points to start with for fine-tuning a pre-trained model, in this case 30,000 training samples, the baseline model performance is already very close to the model trained with complete dataset. The test accuracy is around **92.3%** when compared to **92.9%** from the overall baseline model. Given the difference between the test accuracy and training accuracy is around **6%**, both models seem to be over-fitting, as seen from the almost similar validation and test accuracy. Least confidence sampling method improves the accuracy by **0.1%** over the baseline, but that gain can also be attributed to randomness, while no gain is observed by random sampling and entropy sampling. However, random sampling reduces the variance of the model, difference between train and test accuracy of **4%** compared to **6%** for other methods.

In the low data regime, Table II, baseline test accuracy is 6% lower than the overall baseline, while both of them have similar variance of around 6%. Least confidence sampling improves the test accuracy by 1.5%, while increasing the variance to 10%. Increase in variance can be attributed to two factors:

- **20%** samples are added by annotation, when compared to **3.3%** in high data regime
- almost all of the sampling methods are biased towards the classes on which the model is not doing well. This introduces imbalance in the training data. From Table III we can observe, around **80%** of samples from least confidence and **85%** of samples from entropy sampling are tagged to World and Business News.

Random sampling, Table III shows an interesting pattern. Distribution of the world news and sports news are very close to 25%, as would be expected in the case of balanced dataset, whereas business news is more than twice that of science and technology. The reason behind this is that the AG's news corpus is very US focused, a part of the world dominated by technology companies, and most of the technology companies' news are about mergers and acquisition, which a lot of time has been tagged as technology news, rather than business news. See Table IV for a few examples of the aforementioned type and cases where text is not very clear.

We compared the samples drawn using entropy and least confidence method for similarity. We observed that around 95% and 65% of the samples are common in the high and low data regime, across two sampling methods. This along with test set accuracy indicates, superiority of least confidence sampling over entropy.

VII. ANNOTATION IS HARD

From the Table IV, it is evident why annotation is hard in NLP. It is mostly due to **ambiguity** - first and second examples, and **incomplete information** in text - last two examples. Annotators **might not have the context to the local references and culture**, especially if the data represents different geography than that of the annotator.

For better and faster annotation, the annotation tool should have clear instructions for each of the classes along with proper examples, while providing the option for the user to accommodate their confusion (providing a separate bucket where unsure data points can be tagged). As a possible feature, the tool should have the ability to pull information around the named entity to make the annotator aware of the context so they can tag the data into the appropriate class. The tool

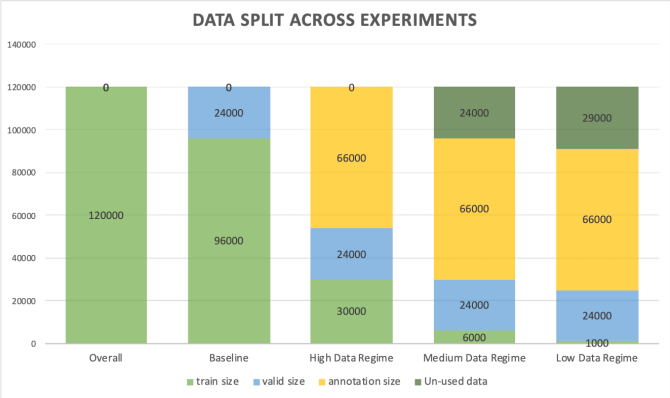

Fig. 1. Data split across various experiments

should also have a feature to avoid showcasing duplicates and junk text, example **#NAME?** occurs ten times in the data, and it was shown for tagging (it's not dealt with in our tool's current state). Our custom annotation tool had clear instructions, examples, ability to undo an annotation and tag a text to none of the class.

VIII. Conclusion

In a transfer learning setting, active learning can be leveraged with as many as 1000 samples. In the small data regime, with 1000 or less samples, least confidence sampling is more data efficient, as it gives the best jump in test accuracy, while increasing variance of the model. In the medium data regime, when test accuracy is close to the baseline, it's best to use random sampling to reduce the variance of the model, while maintaining similar performance (to be confirmed by future study). In the high data regime, with one-fourth of the actual data size, there is almost no gain from active learning, indicating more value could be obtained via model tuning, or data cleaning.

References

[1] Text Data Search - Papers with Code
[2] AG's News Corpus Dataset
[3] AG's News Corpus Dataset, Kaggle
[4] Simple Transformers
[5] Hugging Face Transformers
[6] RoBERTa: A Robustly Optimized BERT Pretraining Approach
[7] Made With ML

TABLE I
ACCURACY FOR HIGH DATA REGIME

Sampling Method	# Train Samples	# Test Samples	# Annotated Samples	Train Accuracy	Valid Accuracy	Test Accuracy
Overall Baseline	96,000	7,600	0	98.6%	92.9%	92.9%
Baseline	30,000	7,600	0	98.7%	93%	92.3%
Random	30,000	7,600	1,000	95.4%	92.1%	91.6%
Entropy	30,000	7,600	1,000	98.4%	92.3%	91.8%
Least Confidence	30,000	7,600	1,00	98.4%	92.6%	92.4%

TABLE II
ACCURACY FOR LOW DATA REGIME

Sampling Method	Train Samples	Test Samples	Annotated Samples	Train Accuracy	Valid Accuracy	Test Accuracy
Overall Baseline	96,000	7,600	0	98.6%	92.9%	92.9%
Baseline	1,000	7,600	0	92.9%	87.4%	86.9%
Random	1,000	7,600	200	98.6%	88.2%	87.8%
Entropy	1,000	7,600	200	98.9%	88.6%	88.4%
Least Confidence	30,000	7,600	200	98.5%	88.8%	88.5%

TABLE III
DISTRIBUTION OF ANNOTATED SAMPLES ACROSS 4 CLASSES FOR LOW DATA REGIME

Class	% Samples: Entropy Based Sampling	% Samples: Least Confidence Sampling	% Samples: Random Sampling
World News	42.1%	42.3%	24.5%
Sports News	3.6%	3.7%	20.5%
Business News	42.6%	36.8%	37.5%
Science / Technology News	11.7%	18.2%	17.5%

TABLE IV
DIFFICULT TO ANNOTATE TEXT

Text	**Actual Class**	**Annotated Class**
Peoplesoft inc. will push ahead with a marketing alliance and other initiatives in the face of a court ruling that raises the chance that oracle corp.'s $7.7 billion hostile takeover bid could succeed, Peoplesoft's chief executive said yesterday	Business	Business
Reuters - Peoplesoft inc. will push ahead with a new marketing alliance and other initiatives in the face of a court ruling that raises the chance that oracle corp.'s 7.7-billion hostile takeover bid could succeed, Peoplesoft's chief executive said on tuesday	Technology	Business
Brussels (Reuters) - Microsoft corp is ready to ask a judge on thursday to suspend penalties imposed on it for violating antitrust law by using the monopoly of its windows operating system to hurt competitors	Technology	Business
The cable television company has taken an aggressive and, critics say, sometimes heavy-handed, approach to protecting its interests in the Maryland county	Technology	Not Sure (In-complete sentence)
coach joins the s p 500, and others stand to benefit from the leather in the weather	Business	Not Sure (In-complete sentence)
#NAME?	World	Not Sure (Junk text)