Robust Active Learning Strategies for Model Variability

José Mena
Eurecat Centre Tecnològic de Catalunya
Universitat de Barcelona
jmenaro17@alumnes.ub.edu

Matthias Gallé
NAVER LABS Europe
matthias.galle@naverlabs.com

Abstract

Active learning methods are useful when a limited budget for data labelling is available. However, the most widely used methods – uncertainty sampling – may suffer from problems derived from an excessive dependence on the model learned during data acquisition. This results in datasets which are not optimal when they are used to train models very different from those used during data creation. In this paper, we link this to the tendency of uncertainty sampling to select outliers and show that other methods that favour selection of representative sampling are more robust to changes in models. We validate this experimentally on four NLP datasets.

1 Introduction

Active learning is a powerful technique to get the best out of an annotation budget for machine learning services in general, and Natural Language Processing ones in particular. It consists of integrating the current model itself in the selection of which data points should be annotated next, which are selected based on an heuristic that is supposed to maximize the improvement of the model. Recent work however [Lowell et al., 2019] has drawn attention to a dangerous side-effect of data-sets created through such models. Because the data selection is model-specific, the best selection for one model might not be the same than the best selection for another one. This becomes an issue in particular when datasets outlive model usage, as has happened in NLP in the last years.

This point was already made 10 years ago. In an influential paper, Settles said [Settles, 2011, Sect 2.6]:

> An important side-effect of active learning that the resulting labeled training set $\mathcal{L}$ is not an i.i.d. sample of the data, but is rather a biased distribution which is implicitly tied to the model used in selecting the queries. Most work in active learning has assumed that the appropriate model class for the task is already known, so this is not generally a problem. However, it can become problematic if we wish to re-use the training data with a model of a different type—which is common when the state of the art advances—or if we do not even know the appropriate model class (or feature set) for the task to begin with

In this overview, Settles cites pre–deep-learning work that made similar arguments and that propose as an alternative the use of ensemble models to obtain data points that improve performance across a variety of models [Baldridge and Osborne, 2004, Lu et al., 2010, Sugiyama and Rubens, 2008].

In this paper, we analyse how this issue affects NLP classification problems, and we show the effectiveness of a much simpler solution. The key insight is that the most used active learning strategy, namely *uncertainty sampling*, tends to select outliers: data points that are far apart from the underling generative distribution $p(x)$. As different models approximate $p(y|x)$ differently, uncertainty sampling tends to pick up extreme elements *with respect to the selection model* that are not necessarily representative. We benchmark this observation empirically on NLP datasets, and show that this phenomenon gets mitigated by using diversity-based acquisition functions. Diversity Sampling methods will find points representative of different parts of the distribution, and not only those close to the decision boundary. When used in combination with uncertainty sampling, diversity enriches the training samples and, therefore, fosters the learning process.

Our contribution consists in benchmarking data selection criteria for their robustness across a variety of deep and linear models. The experiments on 4 datasets show that using active learning strategies less prone to select outliers are more robust to model change.

2 Notation and Related Work

In the present work, we focus on active learning strategies for textual document classification. In order to measure the robustness of selection criteria, we will differentiate between the model that is used to select data points ($\mathcal{A}$, the *acquisition model*) and the model that is used to train on the final data-set ($\mathcal{S}$, the *successor model*) [Lowell et al., 2019]. The acquisition model is used to select different batches of data points to annotate, which we – as is common practice – simulate by retrieving their annotations. In this form, incremental annotated data-sets $\mathcal{D}_A^t$ are created. We assume that at each time-step a fixed number (B) of data points are selected from the full unlabelled data-set $\mathcal{U}$. After T iterations, the final data-set $\mathcal{D}_A^T$ is then used to train the successor model $\mathcal{S}$ and the performance of that model on the unseen test-set $\mathcal{D}_{test}$ is reported.

The active learning process starts with a warm-start dataset, $\mathcal{D}_A^0 = \mathcal{D}_W$, and the batch of new data points at time step t is denoted by $\mathcal{C}_A^t$, using an acquisition function f_{ac}. The full algorithm is shown in Alg. 1.

Algorithm 1 Active learning template algorithm

```
Require: $|\mathcal{U}| \gg |\mathcal{D}| \wedge T > 0$
  $accuracies \leftarrow [\quad]$
  $\mathcal{D}_A^0 = \mathcal{D}_W$
  for $t \in [1 \ldots \mathcal{T}]$ do
    $\mathcal{A} \leftarrow train\_model(\mathcal{D}_A^t)$
    $\mathcal{S} \leftarrow train\_model(\mathcal{D}_A^t)$
    $accuracies[t] \leftarrow get\_accuracies(\mathcal{S}, \mathcal{D}_{test})$
    $select\_indexes \leftarrow select\_items(\mathcal{U}, \mathcal{A}, f_{ac})$
    $\mathcal{C}_A^t \leftarrow \mathcal{U}[select\_indexes]$
    $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{C}_A^t$
    $\mathcal{D}_A^{t+1} \leftarrow \mathcal{D}_A^t \cup \mathcal{C}_A^t$
  end for
```

3 Setting

In our experiment we keep B fixed, and we experiment across 4 datasets ($\mathcal{U}$), 4 acquisition functions (f_{ac}), and 3 acquisition ($\mathcal{A}$) and successor ($\mathcal{S}$) models.

3.1 Datasets

Regarding datasets, we use 4 standard benchmarks, varying across domains and number of classes, presented in Table 1. All the datasets are split into an 80-20 split among train and test samples (for each setting we perform several runs). No special pre-processing of the texts is performed, except for `newsgroup` where documents with less than 10 tokens are removed.

3.2 Acquisition functions

As standard baseline (and comparison point), we use *random* that samples data points from a uniform distribution. The most popular selection criterion is *uncertainty sampling* which uses the probability

Table 1: Datasets used in the experiments

Dataset	**Domain**	**# documents**	**# classes**	**Avg length**
Newsgroups [Lang, 1995]	Newsgroups posts	15,174	20	400
TREC [Li and Roth, 2002]	Questions type classification	5,952	6	48
Subjectivites [Pang and Lee, 2004]	Movie reviews (objective or subjective)	10,000	2	130
Movies [Pang and Lee, 2005]	Movie reviews (positive or negative)	10,662	2	115

of predicted classes. Data points are scored according to the entropy of $p_{\mathcal{A}}(y|x)$, prioritizing higher-entropy points (less certainty).

In addition to this, we study acquisition functions proposed in Zhdanov [2019], which aim to increase diversity in the selection of new items to annotate. On top of *informative* examples, those strategies also aim to select *representative* ones. This is achieved by clustering selected data points and choosing only one representative per cluster: $\beta * B$ ($\beta > 1$) data points are selected using uncertainty sampling and used as input of a clustering algorithm to produce B clusters, each one of which will provide one data point to annotate. Here, we benchmark the following clustering strategies:

- K-means++. The standard k-means algorithm with the recommended initialization from Arthur and Vassilvitskii [2007]. The representative is the point closest to the centre of each cluster.
- Weighted k-means++. Similar to previous set-up, but the distance of a data point to the centroid is weighted by its corresponding uncertainty score.
- k-medoids. The representatives are directly the referent data point of each cluster.

3.3 Models

Model variations comprise 1 linear and 2 deep models:[1]

- Support Vector Machines (SVMs): on top of texts represented via sparse, TF-IDF bag-of-words (BoW) vectors. SVM is implemented using a linear kernel.
- Convolutional Neural Networks (CNNs): the text is represented by a $\ell \times d$ matrix, where ℓ is the sequence length (120) and d is the dimension of the Glove embeddings [Pennington et al., 2014]. For the convolution, we applied filter sizes of 3, 4, and 5, with 128 filters per size.
- Bidirectional Long Short-Term Memory (BiLSTM): words are also represented as Glove embeddings, and the maximum sentence length is set such that it covers 90% of the documents [Lowell et al., 2019].

Hyperparameters of deep models were fine-tuned for each task, resulting in values for learning rates between 1e-4 and 4e-4, a number of epochs that went from 100 to 150 and for all cases a batch size of 128 samples and a dropout rate of 50%.

4 Experiments

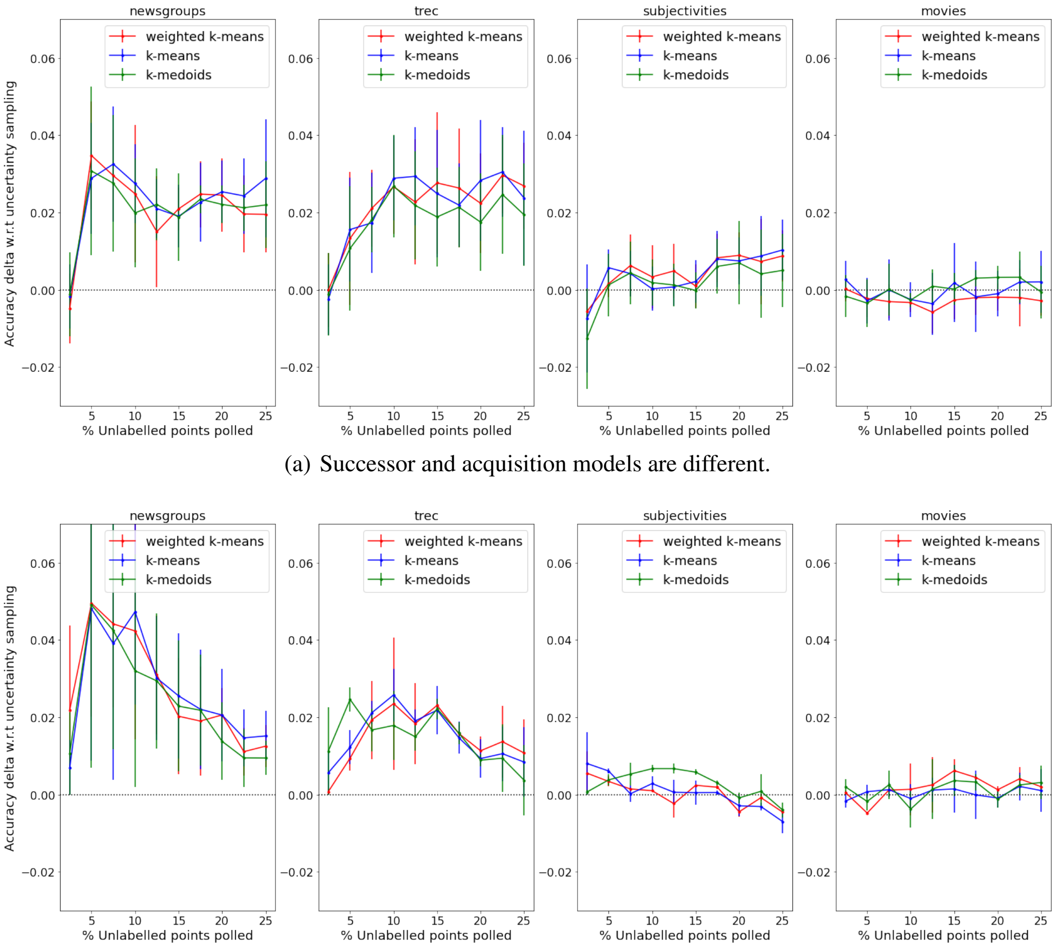
We performed one experiment for each combination of dataset, acquisition function, acquisition model and successor model. For each combination, we performed 10 runs and report the mean and standard deviation of the corresponding accuracy on the test set. The warm-up dataset ($\mathcal{D}^0$) is a random subset, of size 2.5%, each batch B corresponds also to 2.5% of the total data-set and we performed 10 time-steps (T).

In Appendix A we show the accuracies obtained with the different combinations. Here, we focus on a direct comparison with uncertainty sampling, and plot the difference in accuracy with respect of the other strategies with respect to that one. Fig. 1 plots that average difference, for each combination of acquisition model $\rightarrow$ successor model. By averaging differences, the relative performance of each model is taken care of (computing the difference normalizes against that variance). We can conclude from there that in general the cluster-based acquisition functions are more robust to the drop in

[1] implemented following `https://github.com/asiddhant/Active-NLP`

performance due to model change, with 3 absolute points of difference in some cases (newsgroups and trec).

In order to control for the fact that the benchmarked acquisitions functions are just better per se, we also plot that difference when there is **no** change between acquisition and successor model in Fig. 1b. As can be seen in this case the difference with uncertainty sampling gets reduced as more data points are annotated (except for movies), arguably because of the overlap between the datasets $\mathcal{D}^t_{\mathcal{A}}$ increases as t increases. The fact that this does not happen when there is a change in the models seems to indicate that the improvement of the diversity-based methods is maintained even when trained on more data.



(a) Successor and acquisition models are different.

(b) Successor and acquisition models are the same.

Figure 1: Difference of accuracy with respect to uncertainty sampling, measuring the robustness to model shift (a) and their improvement when the model stays the same (b).

5 Bias towards Outlier Selection

Here we investigate one hypothesis for the reason of the better results achieved with diversity-based acquisitions functions. We test if this might be due (or correlated) to the fact that uncertainty sampling is prone to select noisy or out-of-distribution points that fall out of the main data distribution $p(x)$, as approximated by the acquisition model. The relationship between the selection of uncertainty sampling (which focuses on $p(y|x)$) and the *outlier-ness* of a data point (related to $p(x)$) is not straightforward. However, it is common practice to modify acquisition functions to select representative samples Settles [2011].

We investigate this empirically as follows: we start by selecting 2.5% random data points and train a model on this subset. At line 7 of Alg. 1, we enrich the unlabelled pool $\mathcal{U}$ by 10% with data points from another dataset. We perform three runs, where in each run we inject random data points from only one of the three other datasets. Afterwards the 200 data point with the highest score given by uncertainty sampling are selected. If the selection would be random, then 10% of the selected data points would come from the out-of-domain dataset.

Fig 2 plots this proportion for different datasets and models. Each bar is the average of the three runs. The biggest influence on how much uncertainty sampling selects outliers is the dataset, with the biggest proportion for `newsgroup`, where 80% of the selected data points are outliers on average (for BiLSTM). In general, uncertainty sampling is indeed much more sensitive to select such outliers, excepting when using CNN as the base model.

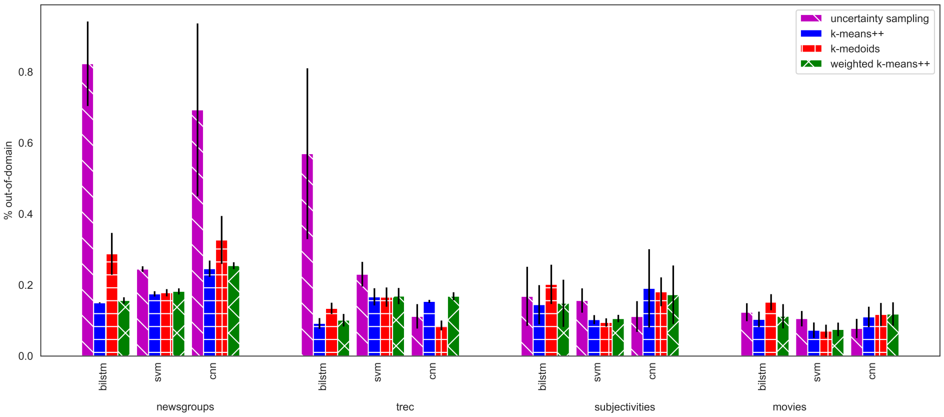


Figure 2: Percentage of selected (using uncertainty sampling) outliers, which were 10% of the pool of unlabelled data points. Each of the four group corresponds to one in-domain dataset, and each of the three subgroup is one base model.

Most interestingly for this paper, there is a correlation between the difference of outlier selection in Fig. 2 and the magnitude of the difference in Fig. 1a. For `trec` and `newsgroups` uncertainty sampling is very biased towards outlier, and indeed, the difference in accuracy in Fig. 1a is the highest; while this difference is much lower for the other two data-sets.

6 Conclusions

We have analysed the sensitivity of active learning strategies against model change in NLP problems. In particular, we have pinpointed the tendency of uncertainty sampling to select outliers concerning the acquisition model: diversity-based models not only improve upon this, but their improvement is correlated to their robustness against outliers. While they sometimes still perform worse than the random baseline (see the Appendix), they do much better than uncertainty sampling.

Acknowledgments and Disclosure of Funding

This work was partially funded through the European Unions Horizon 2020 research and innovation programme under grant agreement No. 786741 (SMOOTH Platform).

References

David Arthur and Sergei Vassilvitskii. K-means qh-: The ad-vantages of careful seeding. In *The 18th Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans*, 2007.

Jason Baldridge and Miles Osborne. Active learning and the total cost of annotation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 9–16, 2004.

Ken Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339, 1995.

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

David Lowell, Zachary C. Lipton, and Byron C. Wallace. 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, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1003. URL `https://www.aclweb.org/anthology/D19-1003`.

Zhenyu Lu, Xindong Wu, and Josh Bongard. Adaptive informative sampling for active learning. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 894–905. SIAM, 2010.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, 2004.

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

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *In EMNLP*, 2014.

Burr Settles. From theories to queries: Active learning in practice. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, pages 1–18, 2011.

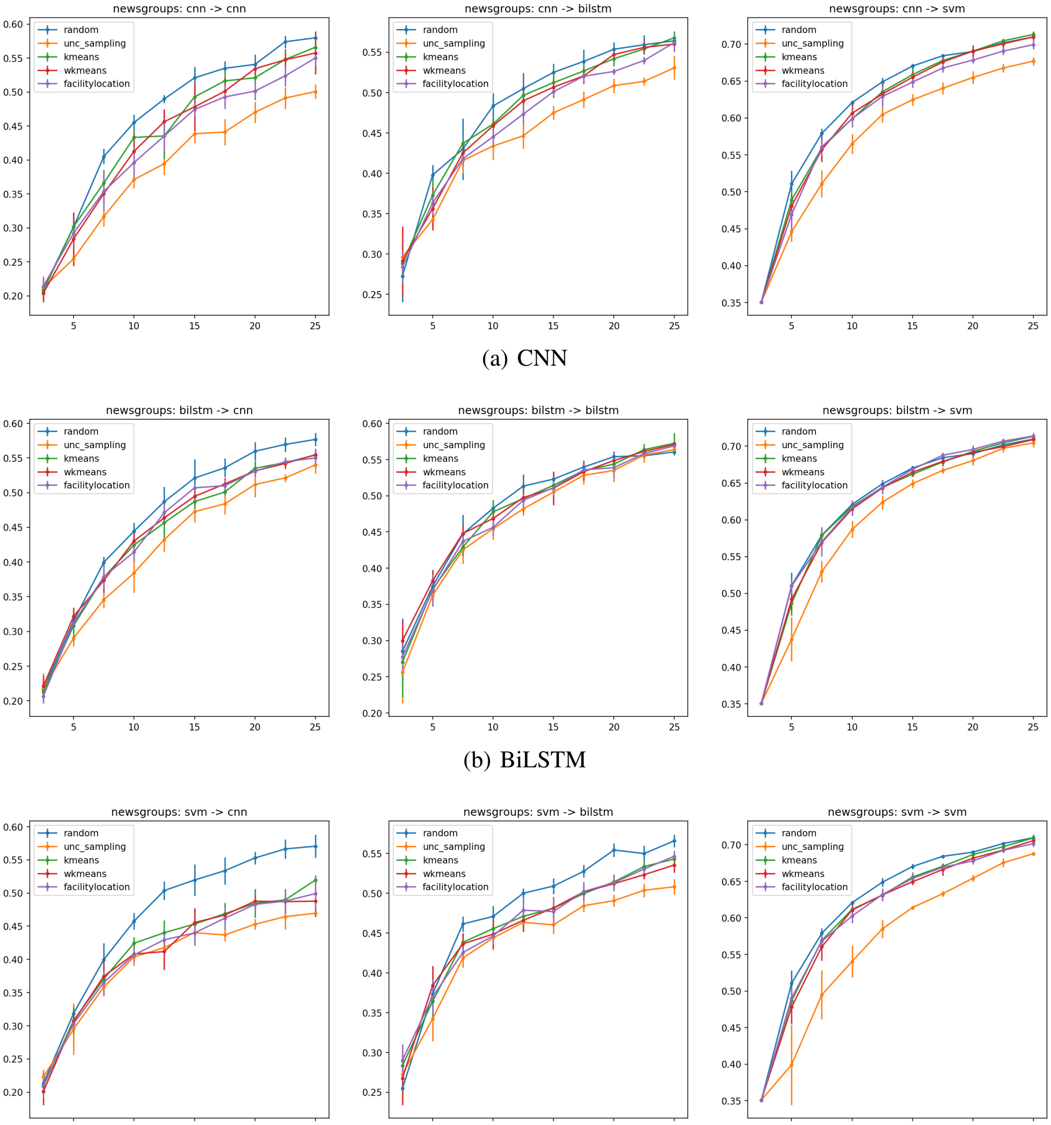
Masashi Sugiyama and Neil Rubens. Active learning with model selection in linear regression. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 518–529. SIAM, 2008.

Fedor Zhdanov. Diverse mini-batch active learning, 2019.

A Appendices

The following tables show the results of comparing the different acquisition functions for active learning analysed in the paper with a non-active learning approach that chooses elements to label randomly.

For each dataset included in the experiments, we plot the results for the different combinations of acquisition and successor models when selecting the different architectures (CNN, BiLSTM and SVM)

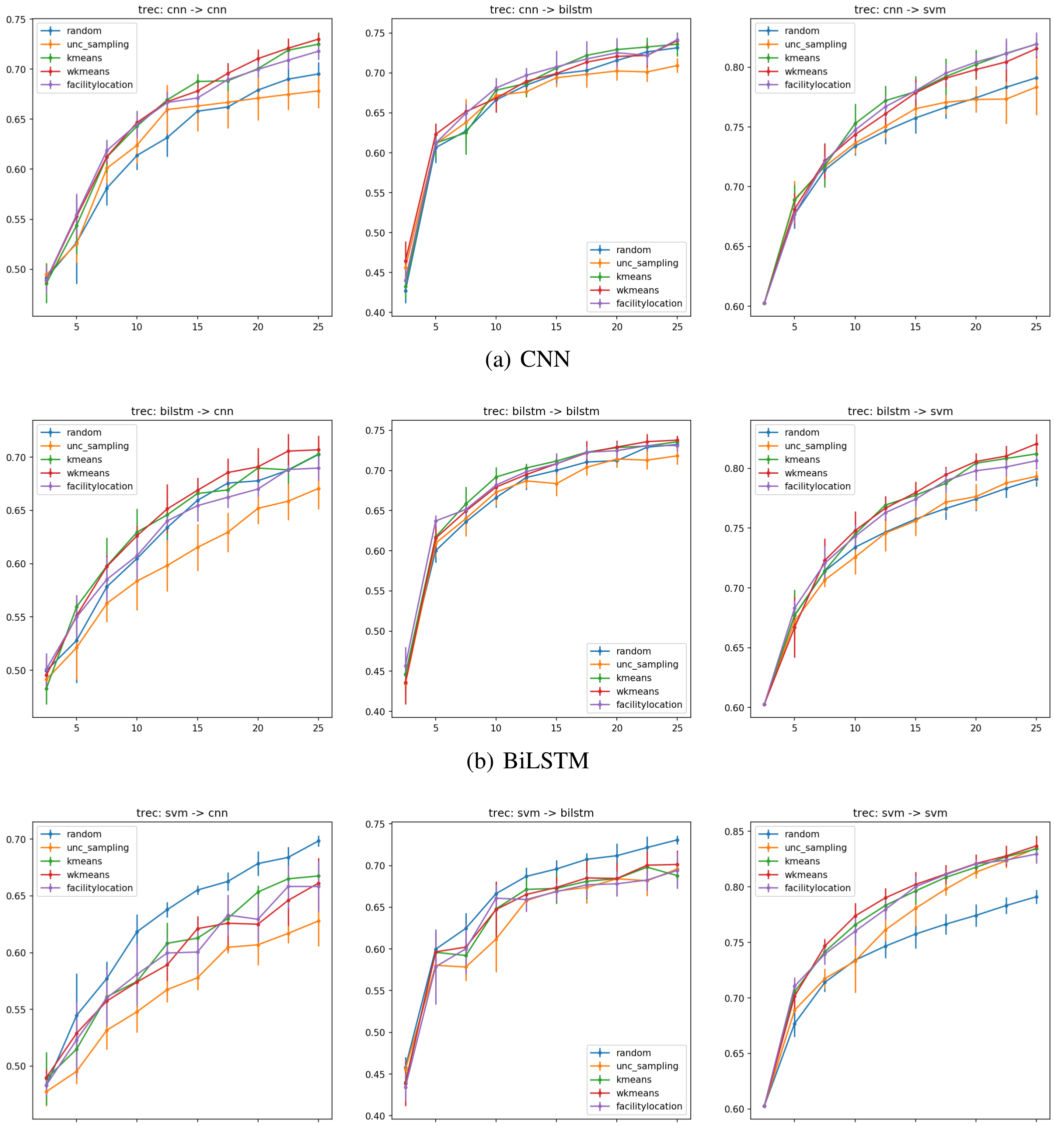


(a) CNN

(b) BiLSTM

(c) SVM

Figure 3: Results on `newsgroup`.

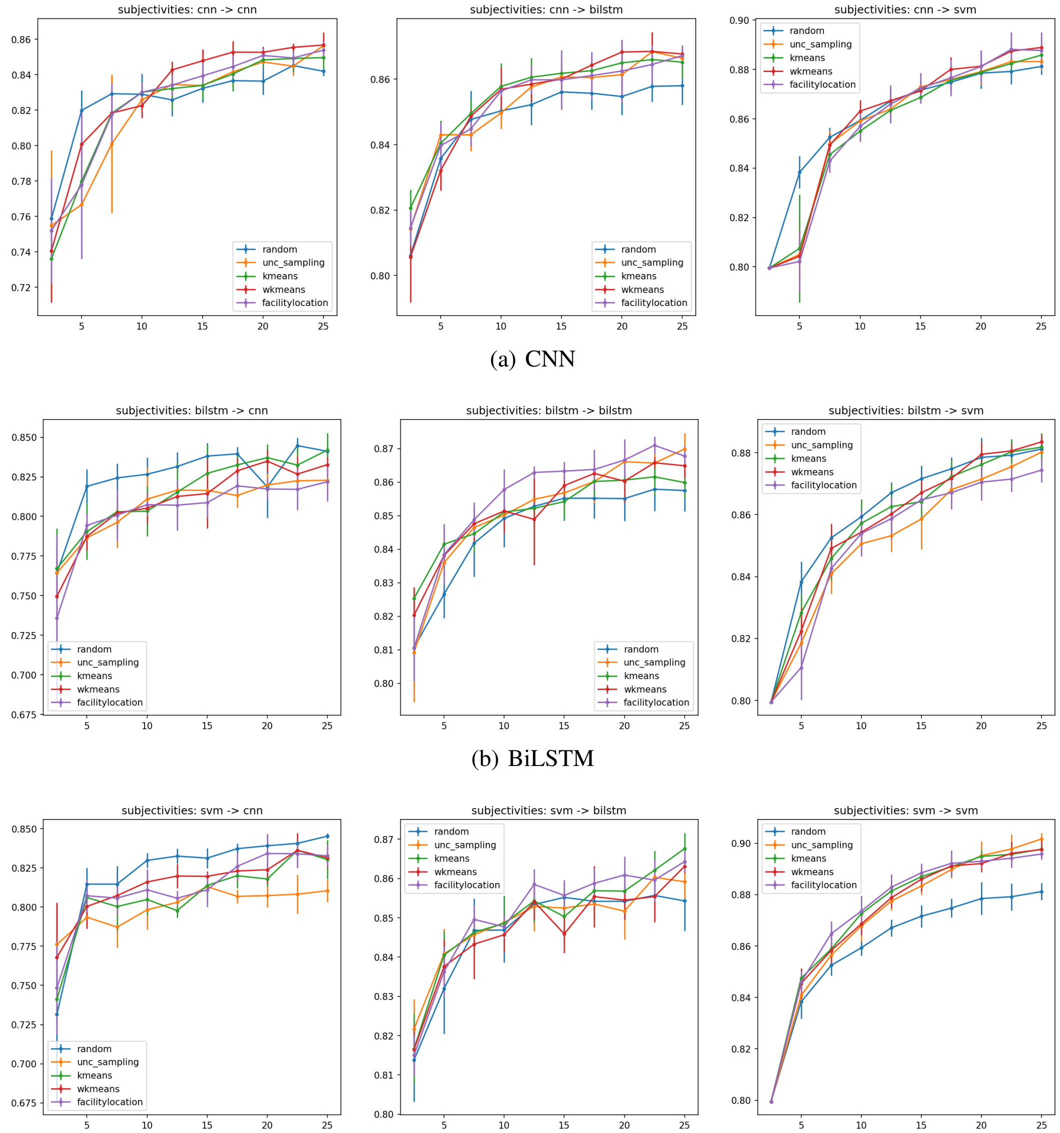


(a) CNN

(b) BiLSTM

(c) SVM

Figure 4: Results on `trec`.

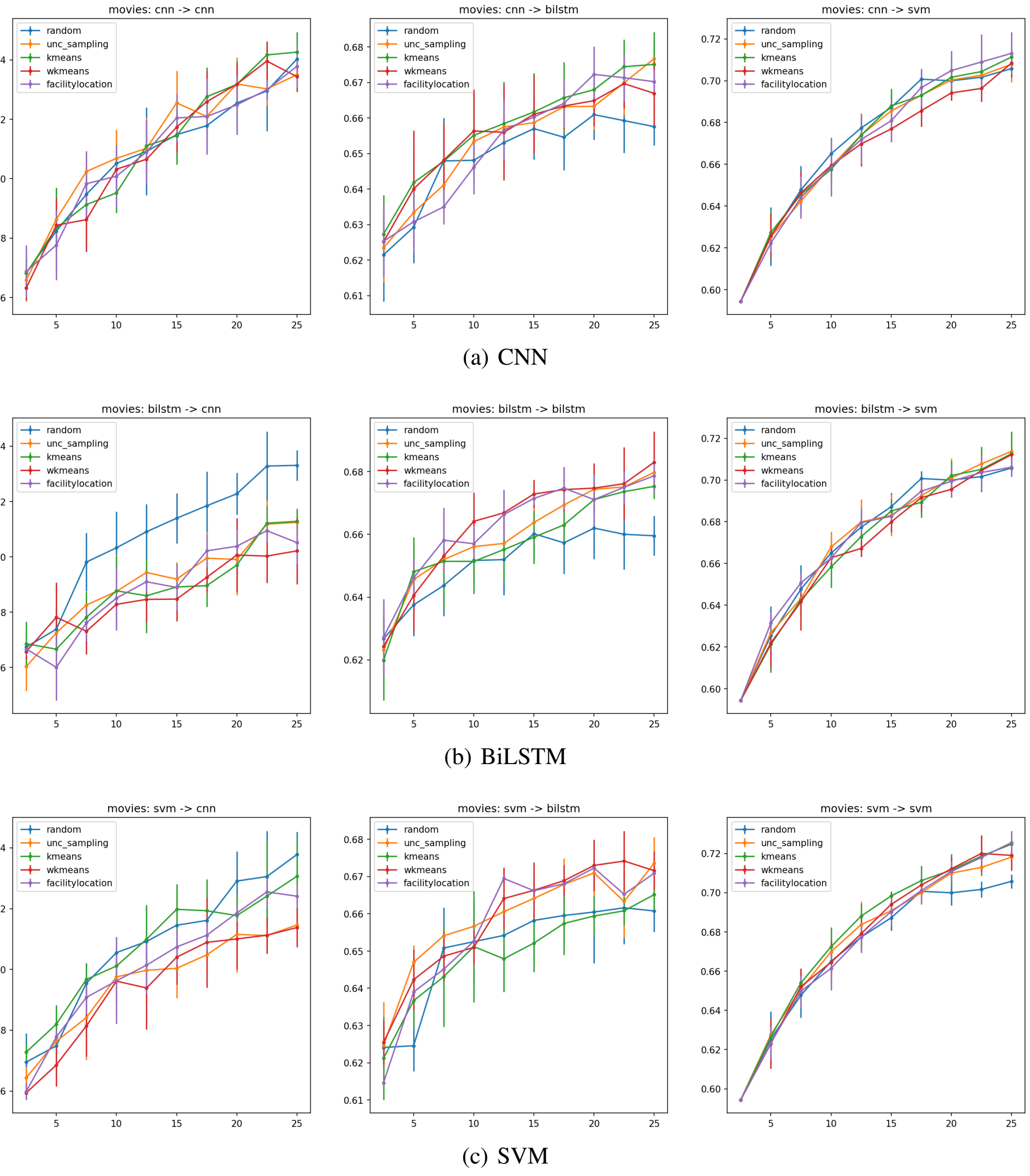


(a) CNN

(b) BiLSTM

(c) SVM

Figure 5: Results on `subjectivity`.



(a) CNN

(b) BiLSTM

(c) SVM

Figure 6: Results on `movies`.