Evaluating Pre-Trained Sentence-BERT with Class Embeddings in Active Learning for Multi-Label Text Classification

Lukas Wertz
University of Stuttgart
lukas.wertz@ims.uni-stuttgart.de

Jasmina Bogojeska
Zurich University of Applied Sciences
bogo@zhaw.ch

Katsiaryna Mirylenka
IBM Research – Zurich
kmi@zurich.ibm.com

Jonas Kuhn
University of Stuttgart
jonas.kuhn@ims.uni-stuttgart.de

Abstract

The Transformer Language Model is a powerful tool that has been shown to excel at various NLP tasks and has become the de-facto standard solution thanks to its versatility. In this study, we employ pre-trained transformer document embeddings in an Active Learning task to group samples with the same labels in the embedding space on domain-specific corpora. We find that the calculated class embeddings are not close to the respective samples and consequently do not partition the embedding space in a meaningful way. In addition, using the class embeddings as an Active Learning strategy yields reduced results compared to all baselines.

1 Introduction

While text classification models have become more and more powerful, the need for sufficient data to train ever growing neural networks is also increasing massively. When dealing with domain-specifc data, such as legal or medical in particular, finding a fitting dataset with detailed annotations can be exceedingly difficult. Creating such a dataset is likely to be a massive undertaking due to the difficult annotation process which often requires domain experts to work through enormous amounts of data. Active Learning serves as a way to speed up this process by selecting informative samples to be annotated. However, Active Learning strategies are often very specific to target domains (Wertz et al., 2022) and strategies tailored specifically for pre-trained transformer language models are often experimental and not thoroughly explored (Zhan et al., 2022).

In this work, we present an Active Learning strategy that employs class embeddings which are generated from pre-trained sentence embeddings to predict the classes of unlabeled samples. While the intuition of the approach is sound, we find that the class embeddings do not generalize from the samples they were calculated on. Our experiment focuses on powerful pre-trained, transformer sentence-embeddings which are prevalent in both research and industrial application. We demonstrate that such embeddings struggle to find good separations between the multi-class, multi-label texts in the training set on two domain-specific datasets. Our work details the class embedding approach, illustrates the reduced performance on two domain-specific, multi-label datasets and analyses the vector space of the samples to gain an understanding of the methods failure.

2 Related Work

The effectiveness of AL for Text Classification has been subject to extensive research (Tong and Koller, 2001), (Goudjil et al., 2018) with specific solutions for deep models (Schröder and Niekler, 2020), (An et al., 2018) and multi-label settings (Reyes et al., 2018) (Yang et al., 2009). Our approach targets Active Learning for Deep Learning which poses new challenges (Schröder and Niekler, 2020) and is still a topic in need of exploration (Ein-Dor et al., 2020). Generating embeddings from words has been performed with trained vector models (Church, 2017) (Pennington et al., 2014) but has been moved to the contextual embedded information within large transformer language models such as BERT (Devlin et al., 2018). Extracting embeddings across word boundaries from BERT can be done in several ways, such as a grid-based approach (Denk and Reisswig, 2019), a "siamese" dual network architecture (Reimers and Gurevych, 2019) or unsupervised techniques (Zhang et al., 2020).

3 Class Embeddings

3.1 Intuition

In any text classification task, the aim is to identify the belonging of a text T to a range of pre-defined classes C. Using pre-trained language models, a

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 366–372
November 20–23, 2022.

text classification model M decides the class $c \epsilon C$ using only the tokenized text as input, leveraging the powerful pre-trained weights of the underlying transformer network as information. We can thus assume that the surface tokens are the critical information that determine, what class T is assigned.

One option to represent text in a continuous vector space is via *embeddings* - vectors that are conditioned to correspond to pieces of text. We convert T into the vector space via embeddings (T_e). Intuitively, one would assume that T_e which belong to the same c are also closer together in the vector space. After all, if c is mainly decided based on the surface tokens, it follows that there should be either syntactical or semantical similarity between two T both belonging to c. While semantical similarity is much harder to capture than the surface realisation of language, current text embedding techniques have shown to also be sensible to word meaning (Wiedemann et al., 2019).

In conclusion, we expect T that belong to the same class to be closer together in a fitting vector space representation because their text should show similarities. Consequently, we assume that if a new text T^* is mapped into the same vector space, it is more likely to belong to the same classes as its neighbours. As such, the centroid of a set of T_e can be used to predict the class of said T^*.

3.2 Active Learning with Class Embeddings

$$C_e = \{mean(T_e) | T \epsilon D \text{ and } T \text{ belongs to } c\} \quad (1)$$

Active Learning is a cyclic, supervised learning mechanism that seeks to reduce annotation effort by strategically selecting informative samples to be labeled by a human annotator and then given to the model for training. Given an annotated training set D and an unlabeled set U, the main loop of Active Learning can be summarized in three repeating steps:

1. Train classification model M on available data D.
2. Select informative samples from U and pass them to the annotator.
3. Annotate the samples and add them to D.

Given an annotated set D, our approach calculates **Class Embeddings** C_e for each class c by first collecting all T that belong to c and then using an embedding technique to map T into the vector space. The corresponding $c_e \epsilon C_e$ are determined by calculating the centroid of all T_e belonging to c (Equation (1)).

	train	dev	test	Macro F1
eurlex	10.294	1.901	1.905	0.93
arXiv	13.174	13.414	13.131	0.79

Table 1: Split sizes and Macro F1 on the full *eurlex* and *arXiv* datasets.

In the Active Learning setting, we calculate C_e given the current D and then select k samples which are close to the c_e of classes that are less frequent in the training set. The idea is, that finding samples of less represented classes will improve classifier accuracy on that class and consequently, will improve Macro F1. We update and evaluate M after k samples have been selected and repeat this process until an annotation budget is exhausted. The full procedure is detailed in Algorithm 1.

Algorithm 1 Active Learning with Class Embeddings

```
procedure CE(labeled set D, unlabeled set U,
model M, budget b, sample size k)
    while budget > 0 do
        train M on D
        C_e ← Class Embeddings on D
        k* ← k
        while k* > 0 do
            c_min ← least frequent class in D
            T ← TϵU, T closest to c_e of c_min
            annotate T
            D ← D ∪ T
            k* ← k* − 1
            b ← b − 1
```

4 Experiment

4.1 Datasets

We use modified versions of the Eurlex57K (referred to as *eurlex*) (Chalkidis et al., 2019) corpus containing excerpts from European law as well as a collection of abstracts from scientific publication site *arXiv* (https://www.kaggle.com/Cornell-University/arxiv). Both datasets are annotated with several hundred classes and are intended for large-scale, multi-label text classification, meaning that a sample can belong to any number of classes instead of only one. We reduce the number of classes to 5 frequent and 5 rare labels to create a reduced version of the corpus, keeping the multi-label nature intact. Macro F1 when using the full dataset is found in Table 1.

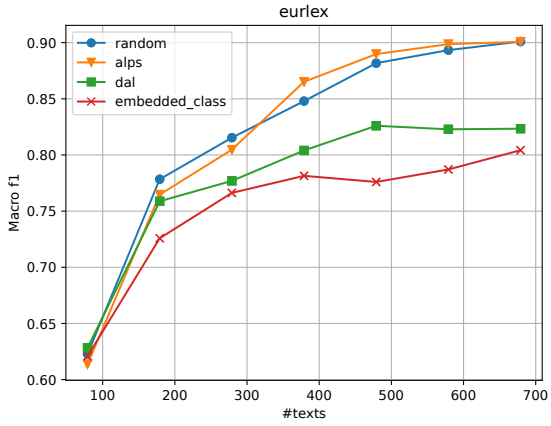

Figure 1: Macro F1 on the *eurlex* dataset of Active Learning for training set sizes 100 to 600 samples compared to random selection and two Active Learning baselines.

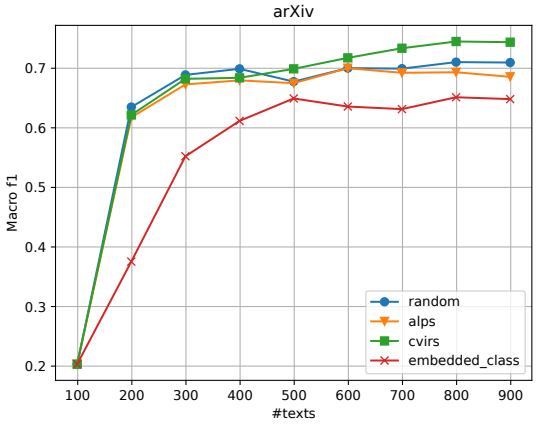

Figure 2: Macro F1 on the *arXiv* dataset of Active Learning for training set sizes 100 to 600 samples compared to random selection and two Active Learning baselines.

4.2 Setup

We use BERT (Devlin et al., 2018)* for text classification with a single, feed-forward output layer. We train the model for 15 epochs with early stopping, a batch size of 16 and an adaptive learning rate (ADAM). We evaluate all experiments using the multi-class measures Macro F1† (averaging F1 for each class, thus, treating each class as equally important, which is beneficial in the unbalanced class settings).

For document embeddings, we employ pre-trained Sentence-Bert (Reimers and Gurevych, 2019) embeddings‡ which maps a document into a 380 element vector.

We simulate Active Learning by using a subset of the corpus as "labeled" set and reserving the rest as the "unlabeled" set, using the oracle annotations once a sample is queried from the "unlabeled" set. We start with a labeled set of 100 randomly selected samples and query 50 samples in each Active Learning step until the annotation budget of 600 samples is exhausted.

All experiments are run on a NVIDIA RTX 6000 GPU.

4.3 Results

Figures 1 and 2 show the results of Active Learning on the *eurlex* and *arXiv* datasets respectively. We compare the class embedding approach (Section 3.2) against three Active Learning baselines (DAL - (Gissin and Shalev-Shwartz, 2019), ALPS - (Yuan et al., 2020), CVIRS - (Reyes et al., 2018)) as well as Active Learning by random sampling. Out of the Active Learning strategies, we report the two best performing approaches for each dataset. We find that the class embeddings perform significantly worse than all baselines by a margin of up to 0.15 compared to random selection. Class Embeddings appear to hinder the Active Learning process as they even perform worse than Active Learning strategies which already have reduced performance compared to random selection, i.e. the *DAL* baseline on the *eurlex* dataset.

5 Analysis

5.1 Proximity to unlabeled samples

One important assumption presented in Section 3.1 is, that an unlabeled§ sample $T^* \epsilon U$ will be close in the embedding space to the class embeddings $c_e \epsilon C_e$ of the classes $c \epsilon C$ it belongs to. We test this assumption by analysing how many T^* that belong to c are actually closest to the corresponding class embedding by querying the closest 100 T^* for every c_e. Table 2 shows, that on the *eurlex* dataset for a small labeled set with 100 samples, almost no T^* are near a c_e of a class they belong to. We also see that this is not an effect of the labeled set being too small as increases in the size of D (even to around 50% of the full training set) do not

*Using the "bert-base-uncased" model from *huggingface* https://huggingface.co

†We also evaluated Micro F1 but found that the two behaved similarly.

‡Using the "all-mpnet-base-v2" downloadable from https://www.sbert.net

§Here, *unlabeled* simply denotes that the sample does not come from the training set of the model (Section 4.2).

size of D	class 1	class 2	class 3	class 4	class 5	class 6	class 7	class 8	class 9	class 10
100	1	0	0	0	0	0	7	2	0	0
200	2	0	0	0	0	0	7	2	0	0
500	1	0	0	0	0	0	7	2	0	0
1500	2	0	0	0	0	0	7	1	0	0

Table 2: Number of samples in the unlabeled set U of the **eurlex** dataset with class *j* found within the closest 100 samples of the centroid of class *j* using pre-trained Sentence-BERT. We experiment with varying sizes of the labeled set D.

significantly change the results. Effectively, this means that the computed c_e are not close to new samples of the same class and that our assumption is incorrect. This observation holds for the *arxiv* dataset as well. (See Appendix for the full results table).

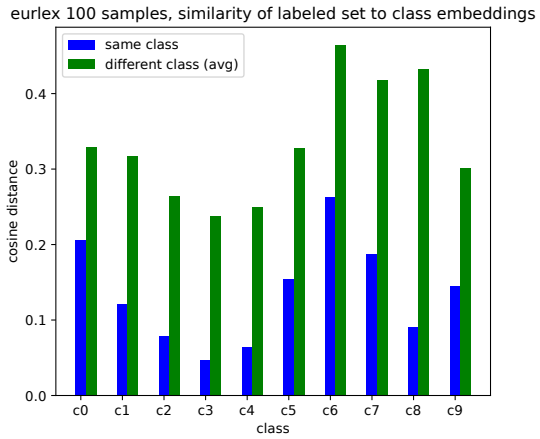

Figure 3: Average cosine distance between labeled samples and corresponding class embedding of the same class (blue, left) and averaged class embeddings of all other classes (green, right).

5.2 Examination of the labeled set

One explanation for the behaviour on unlabeled samples is, that the class embeddings are not well-positioned. For example, when calculating C_e we do not account for outliers which might cause a shift in the centroid. Alternatively, class embeddings might all be very close to each other, resulting in a partitioning that is not very meaningful. We run a sanity check in Figure 3 and Figure 4 and look at the average distance between samples in the labeled set $T \epsilon D$ and the computed class embeddings for a size of 100 samples[¶]. We find that on average, samples are closer to the c_e of classes

[¶]We also experiment with higher numbers but find no significant differences.

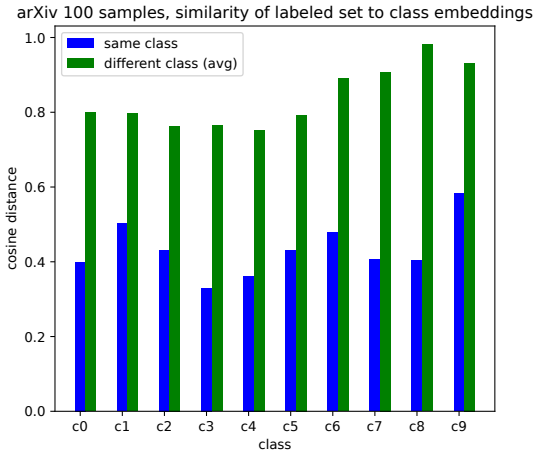

Figure 4: Average cosine distance between labeled samples and corresponding class embedding of the same class (blue, left) and averaged class embeddings of all other classes (green, right).

they belong to by a margin of around 0.2 on the *eurlex* dataset and 0.4 on the *arXiv* dataset. Due to the multi-label nature of the datasets we expect certain overlap between classes. Overall, Figures 3 and 4 seem to indicate a good positioning of the class embeddings, which means that the training set samples are in fact found in the proximity of corresponding class embeddings. Figures 5 and 6 show the result of a Principal Component Analysis (PCA) on the two datasets respectively. We find that while there are some clusters, overall there is no clear separation of classes. This could be an indication, that the sentence-BERT embeddings (see Section 4.2) are too large or too diverse to effectively decompose into 2 dimensions. However, it is also possible that even in the high-dimensional space, separation of the different classes is already difficult.

On the *eurlex* dataset, Figure 3 confirms this suspicion somewhat since the distance margins are narrow overall. We find that for many classes, observations hold between Figure 3 and Figure 5.

For example, samples belonging to class 2 have a are generally very close to their corresponding class embedding while Figure 5 also shows a narrow cluster of class 2 samples. However, for some samples we observe conflicting information from the two Figures, for example class 3, which has the least average distance in Figure 3 but is very spaced out in the PCA in Figure 5.

In general, the analysis of the *arxiv* dataset in Figures 4 and 6 leads to analogous conclusions. The main difference is that while the average distances in Figure 4 are twice as long as for the *eurlex* dataset, the samples in Figure 6 seem even more clustered around a central point. In general, most of the centroids are very close together in the reduced space, making clear separation of classes difficult. Overall, we can conclude that the class embeddings provide only limited grouping for the dataset they were calculated on.

In addition, we find that the labels have semantic overlap to each other. In the *arXiv* dataset, frequent labels deal with various areas of Physics, while rare labels deal with Computer Science and Informatics. On the *eurlex* dataset, frequent labels deal with Fruit, import and export while rare labels are more diverse. (Full Table is found in the appendix). This could explain the proximity of centroids in the PCA analysis, especially for the *arxiv* dataset in Figure 6. On the *eurlex* dataset in Figure 5 however, centroids of different topics, e.g. *Gaming* (centroid 9) and *Export Refund* (centroid 1) are close to each other.

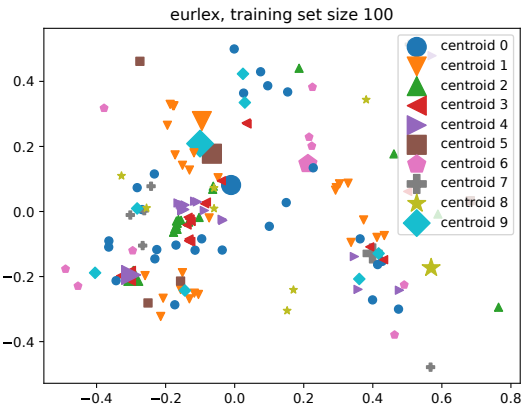

Figure 5: PCA with 2 components of the class embeddings and embedded samples in the training set with 100 samples. Shapes of the data points indicate class (samples with multiple classes are plotted multiple times) and enlarged data points mark centroids (i.e. class embeddings).

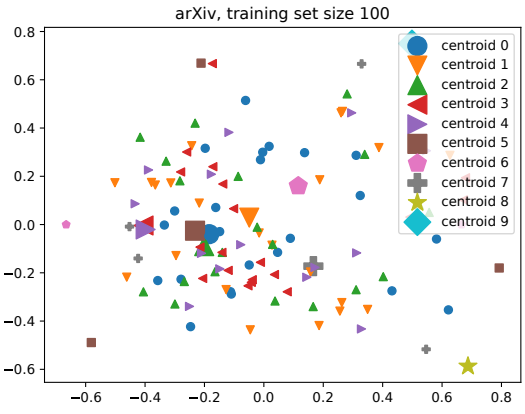

Figure 6: PCA with 2 components of the class embeddings and embedded samples in the training set with 100 samples. Shapes of the data points indicate class (samples with multiple classes are plotted multiple times) and enlarged data points mark centroids (i.e. class embeddings).

6 Conclusion & Future Work

We present Class Embeddings, which hinder the Active Learning (Section 4.3) since the classes of new samples can not be correctly predicted (Section 5.1). Despite reasonable assumptions about the effectiveness of pre-trained embeddings (Section 3.1) we find that class embeddings are not meaningful representatives of the dataset classes and that their ability to partition the dataset is limited (5.2). We encourage experimenting with this approach, as it is relatively inexpensive to compute. In addition to using common heuristics with BERT, such as averaging the word embeddings, fine-tuning the sentence-embeddings on the dataset might make a difference and result in higher quality Class Embeddings. Also, testing the approach on different datasets is crucial - in our work, improving upon random selection is difficult even for sophisticated Active Learning strategies. Finally, we would like to motivate more application-oriented research (e.g. Information Retrieval, Semantic Similarity rankings etc...) into the inner workings of pre-trained contextual embeddings in order to improve understanding of the information they encode.

Acknowledgments

This work was funded and supported by IBM.

References

Bang An, Wenjun Wu, and Huimin Han. 2018. Deep active learning for text classification. In *Proceedings of the 2nd International Conference on Vision, Image and Signal Processing*, pages 1–6.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy. Association for Computational Linguistics.

Kenneth Ward Church. 2017. Word2vec. *Natural Language Engineering*, 23(1):155–162.

Timo I Denk and Christian Reisswig. 2019. Bertgrid: Contextualized embedding for 2d document representation and understanding. *arXiv preprint arXiv:1909.04948*.

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

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Daniel Gissin and Shai Shalev-Shwartz. 2019. Discriminative active learning. *arXiv preprint arXiv:1907.06347*.

Mohamed Goudjil, Mouloud Koudil, Mouldi Bedda, and Noureddine Ghoggali. 2018. A novel active learning method using svm for text classification. *International Journal of Automation and Computing*, 15(3):290–298.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Oscar Reyes, Carlos Morell, and Sebastián Ventura. 2018. Effective active learning strategy for multi-label learning. *Neurocomputing*, 273:494–508.

Christopher Schröder and Andreas Niekler. 2020. A survey of active learning for text classification using deep neural networks. *arXiv preprint arXiv:2008.07267*.

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

Lukas Wertz, Katsiaryna Mirylenka, Jonas Kuhn, and Jasmina Bogojeska. 2022. Investigating active learning sampling strategies for extreme multi label text classification. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4597–4605, Marseille, France. European Language Resources Association.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. *ArXiv*, abs/1909.10430.

Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. 2009. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 917–926.

Michelle Yuan, Hsuan-Tien Lin, and Jordan L. Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. *CoRR*, abs/2010.09535.

Xueying Zhan, Qingzhong Wang, Kuan-hao Huang, Haoyi Xiong, Dejing Dou, and Antoni B. Chan. 2022. A comparative survey of deep active learning.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1601–1610, Online. Association for Computational Linguistics.

Appendix

size of D	class 1	class 2	class 3	class 4	class 5	class 6	class 7	class 8	class 9	class 10
100	1	0	0	0	0	0	7	2	0	0
200	2	0	0	0	0	0	7	2	0	0
500	1	0	0	0	0	0	7	2	0	0
1500	2	0	0	0	0	0	7	1	0	0

Table 3: Number of samples in the unlabeled set U of the **arXiv** dataset with class j found within the closest 100 samples of the centroid of class j using pre-trained Sentence-BERT. We experiment with varying sizes of the labeled set D.

	arXiv	*eurlex*
class 1	High-Energy-Physics	import
class 2	Statistical Mechanics	export refund
class 3	Quantum Physics	Pip Fruit
class 4	Superconductivty	Fruit Vegetable
class 5	Strongly Correlated Electrons	Citrus Fruit
class 6	Atomic and Molecular Clusters	Quantitative Restriction
class 7	Network Architecture	Germany
class 8	Formal Languages	Portugal
class 9	Human Computer Interaction	Ship's Flag
class 10	Other Computer Science	Gaming

Table 4: Descriptions of labels used in both datasets.Frequent labels are above center line, rare labels are below center line.

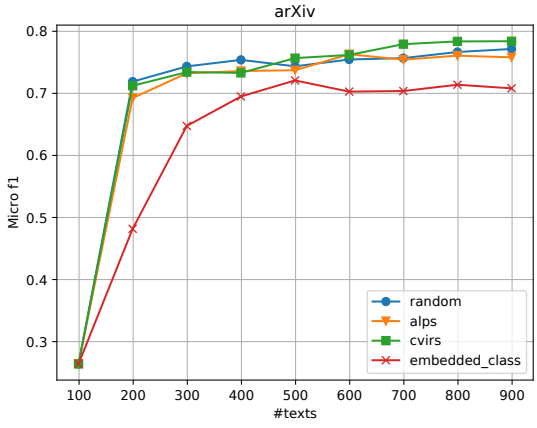

Figure 7: Micro F1 on the arXiv dataset.

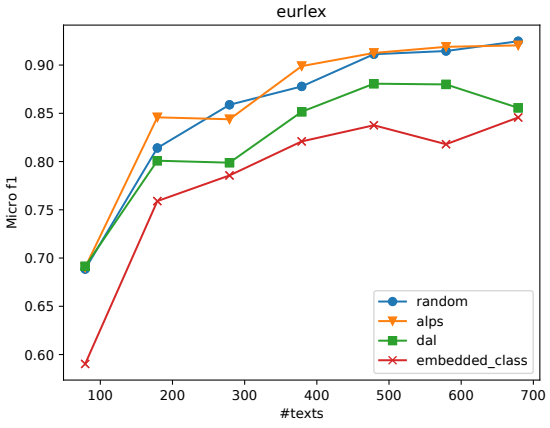

Figure 8: Micro F1 on the arXiv dataset.