Investigating Active Learning Sampling Strategies for Extreme Multi Label Text Classification

Motivation

Most data in the world is unlabeled. Domain-Specific, extremely high label datasets are critically expensive to annotate. We analyse Active Learning (**AL**) strategies for a Extreme Multi-Label Text Classification (**XMTC**) task.

Contribution

We contribute:

- Evaluation of AL strategies on XMTC datasets
- Simple AL strategy that can outperform more expensive approaches
- Tradeoff between increase in F1 and computational cost
- More effective model architecture using CNN on top of BERT
- Multi-Label evaluation with variable threshold

Datasets

	size			number of	classes per	texts
	train	test	dev	classes	text (avg)	per class (avg)
EurLex	44,689	5,954	5,963	739	4.38	265.32
arXiv	239,347	29,174	26,309	113	1.68	4385.46
NYT	22,991	10,941	2,554	303	2.62	315.52
RCV1	20,816	781,265	2,333	100	1.53	12321.26
Yelp	100,911	48,272	11,287	580	2.3	401.68
AGNews	112,400	7,600	7,600	4	0.75	28115.0
Toxic	102,124	25,532	31,915	6	0.22	3702.83

Table: Sizes and class statistics for all datasets. Classes per text and texts per class are averaged over all texts of the respective dataset.

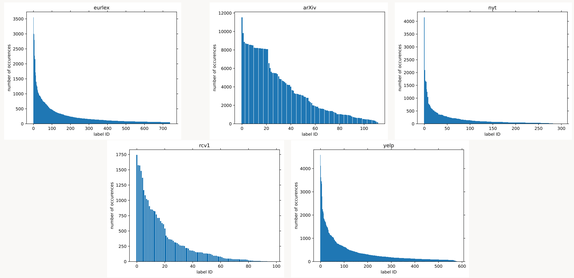

Figure: Class distribution of extreme multi-label datasets.

Acknowledgment

This work was funded by IBM as part of the AI Horizons Network.

Classification Model

- Pre-Trained BERT Language Model
- Convolutional Neural Network as classification output

Computational Efficiency

strategy	computation time for 100 batches (in seconds)
subword	3.80 (computed once)
ALPS	50
DAL	56
CVIRS	123

Evaluation

MicroMacro F1

$$\text{F1}(y, y^*) = 2 * \frac{precision * recall}{precision + recall} \quad (1)$$

$$\text{micro-precision}(y, y^*) = \frac{tp}{tp + fp} \quad (2)$$

$$\text{micro-recall}(y, y^*) = \frac{tp}{tp + fn} \quad (3)$$

$$\text{macro-precision}(y, y^*) = \frac{1}{|C|} \sum_{c \epsilon C} \frac{tp_c}{tp_c + fp_c} \quad (4)$$

$$\text{macro-precision}(y, y^*) = \frac{1}{|C|} \sum_{c \epsilon C} \frac{tp_c}{tp_c + fn_c} \quad (5)$$

Variable Threshold

- Evaluate Multi-Label Predictions with threshold τ
- For a wide range of τ, get F1 from validation set
- Choose τ that gives the best validation F1 (Macro F1) and evaluate on the test set

AL Strategies

ALPS (Yuan et. al. 2020)

- Calculate surprisal embedding -> Vector of BERT language modeling uncertainty
- Cluster embeddings, select closest texts to n cluster centers

CVIRS (Reyes et. al. 2018)

- Difference between predictions on labeled and unlabeled set
- Classification uncertainty ranking for each label

DAL (Gissin et. al., 2019)

- Separate Classifier decides if text comes from labeled or unlabeled set
- Focus on data which comes from the unlabeled set

Subword

- Unkown words in the text are split into subword-units by BERT-tokenizer
- Select texts which have the highest number of subword units

Experimental Results

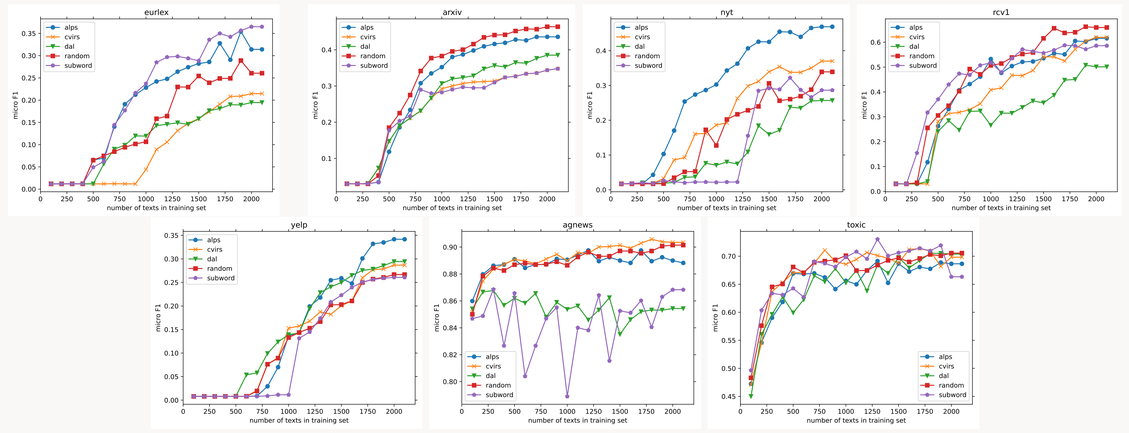

Figure: Micro $F1$ results across all datasets. The x-axis describes the number of texts used to train the classifier while the y-axis shows the resulting micro $F1$.

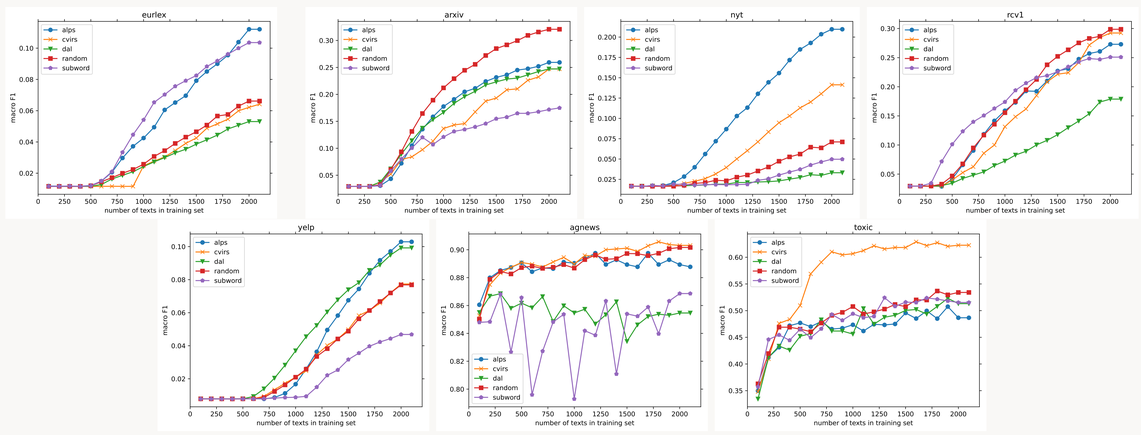

Figure: Macro F1 results across all datasets. The x-axis describes the number of texts used to train the classifier while the y-axis shows the resulting Macro F1.

Lukas Fromme*†, Katsiaryna Mirylenka†, Jonas Kuhn*, Jasmina Bogojeska†

*Institute of Natural Language Processing (IMS) - University of Stuttgart, †IBM Research - Zurich

presented by Lukas Fromme