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 (XMTCL) task.

Contribution
We contribute:
- Evaluation of AL strategies on XMTCL 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

<table>
<thead>
<tr>
<th>dataset</th>
<th>size</th>
<th>number of classes</th>
<th>classes per text (avg)</th>
<th>per class (avg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>eurlex</td>
<td>5,974</td>
<td>250</td>
<td>200.0</td>
<td>25.0</td>
</tr>
<tr>
<td>arXiv</td>
<td>5,974</td>
<td>250</td>
<td>200.0</td>
<td>25.0</td>
</tr>
<tr>
<td>NYT</td>
<td>10,941</td>
<td>200</td>
<td>50.0</td>
<td>25.0</td>
</tr>
<tr>
<td>RCV1</td>
<td>20,816</td>
<td>200</td>
<td>100.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Yelp</td>
<td>48,272</td>
<td>1,000</td>
<td>1,000.0</td>
<td>1,000.0</td>
</tr>
<tr>
<td>AGNews</td>
<td>5,600</td>
<td>500</td>
<td>400.0</td>
<td>400.0</td>
</tr>
<tr>
<td>Toxic</td>
<td>25,532</td>
<td>200</td>
<td>125.0</td>
<td>125.0</td>
</tr>
</tbody>
</table>

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.

Evaluation

MicroMacro F1

\[
F_1(y, y^*) = \frac{\text{precision} + \text{recall}}{2}, \quad \text{macro-precision} = \frac{\sum_i \text{tp}_i}{\sum_i \text{tp}_i + \text{fp}_i}, \quad \text{micro-precision} = \frac{\sum_i \text{tp}_i}{\sum_i \text{tp}_i + \text{fp}_i}
\]

Variable Threshold

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

Experimental Results

AL Strategies

- ALPS (Yuan et. al. 2020)
  - Calculate surprisal embedding \( \rightarrow \) 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
- Unknown words in the text are split into subword-units by BERT-tokenizer
- Select texts which have the highest number of subword units

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.

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

Lukas Fromme*, Katsiaryna Mirylenka†, Jonas Kuhn*, Jasmina Bogojleka†
*Institute of Natural Language Processing (IMS) - University of Stuttgart, †IBM Research - Zurich
presented by Lukas Fromme