**Model Summary**

**Coleridge Initiative – Show US the Data**

|  |  |
| --- | --- |
| Team Name | Hayk Sargsyan |
| Private Leaderboard Score | 0.478 |
| Private Leaderboard Place | 7 |
| Name | Hayk Sargsyan |
| Affiliation | EPha.ch |
| Kaggle Account | https://www.kaggle.com/sarhayk |

**1. Background**

I am a Physicist by education, hold a PhD degree in theoretical physics from University of Zurich. Since 2018 I am working as a Data Scientist at Epha.ch and, recently, was promoted to a CTO position.

In the company, we perform a fare amount of information extraction from medical publications, for example, we extract drug-drug interactions from medical texts. In this respect, I had some prior knowledge, however, since the competition was very unique, it is hard to tell how much this prior experience helped me.

I constantly check the Kaggle competitions. Although, due to the lack of time, I do not manage to participate in many of them, I still check what people are doing and try to learn from it. This competition was different, though, in that I found it very interesting and challenging. Most importantly, it was very distinctive and unclear what kind of strategy might be the best. I started thinking of original ways to tackle the problem and decided to participate in the competition in order to check whether these approaches are competitive to other’s approaches. Unfortunately, I still did not have enough time to finalize everything that I had in mind due to my everyday job responsibilities. I worked pretty chaotic on the competition, therefore it is hard to judge how much time I spent on it, but 10-15 days of full-time work would be a good approximation.

**2. Approach**

Judging from the competition discussion pages it seemed that most of my fellow colleagues were opting for named entity recognition (NER) approaches. I was doubtful that the NER would generalize well, therefore, I started thinking of ways to tackle the problem through text classification. For that, I needed a way of generating dataset candidates. I accomplished it by noticing that most of the datasets in the training set appear with their abbreviations. Therefore I tried to extract abbreviation-definition pairs from the texts and then classified them. In particular, I used the **Schwartz-Hearst1** algorithm for that, with some refinements from my side. For classification I used the **Spacy** NLP tool with the **distillBert** transformer model from **Hugging Face**.

**3. Training Methods**

I did not use any external data for training. As positive examples I used the available dataset labels provided in the training set, while for generating the negative (non-dataset) examples I randomly sampled from the remaining candidates. In the latter, I only filtered those candidates that had a high chance of being datasets. For example, candidate groups that were studies, surveys, programs, etc.. The reason for this was to reduce the noice in the training set by minimizing the share of wrongly labelled negative examples. I believe this approach helped the model to discriminate between positive/negative examples better, which ultimately led to better generalization. I only used one model – no ensembling.

**4. Interesting Findings**

I think the most important trick that helped me to get relatively good results was the abbreviation-definition extraction step with the consecutive grouping of the candidates into groups, similar to “dataset\_title”-”dataset\_label” relationship in the training set.

**5. Model Execution Time**

To train the model end-to-end (from raw data to model checkpoints) takes about 1.5 hrs, using one NVIDIA GTX 1080 TI GPU. The prediction phase (from raw data to labels per publication) takes about 1.5 hrs for around 15000 publications.

[[1]](#footnote-1)

1. # A SIMPLE ALGORITHM FOR IDENTIFYING ABBREVIATION DEFINITIONS IN BIOMEDICAL TEXT, Schwartz, Hearst, 2002 DO - 10.1142/9789812776303\_0042

   [↑](#footnote-ref-1)