CS 6120: Comparing Models for Multilabel Classification of PubMed Article Abstracts

Nithish Bhat and **Ashish Thomas** and **Jamie Tjia**
Northeastern University
Boston, MA, United States
{bhat.nithi, thomas.ash, tjia.j}@northeastern.edu

Abstract

This is a template for your project report. This template is used for submissions to the ACL conference. You can see the original files here (https://github.com/acl-org/acl-style-files/tree/master/latex) which have additional instructions on how to use this template.

1 Introduction

What is this project about? Goals, motivation, etc...

We want to train a transformer model BERT to handle multi label classification for biomedical research articles. ....

2 Background/Related Work

Describe any background or related work (Peng et al., 2021)

3 Data

Describe the datasets and any preprocessing, cleaning that you may have done.

PubMed MultiLabel Text Classification Dataset MeSH contains information on 50,000 articles from PubMed. This information includes 16 MB of data on the article titles, abstracts, and Medical Subject Heading (MeSH) labels.

MeSH labels are arranged into a nested hierarchy with 16 root labels. Each article has up to 13 root labels, with a median root label count of six. Two labels—Humanities [K] and Publication Characteristics [V]—have been ignored due to their infrequency in the available data. Across 50,000 sample articles, exactly one sample has label K and none have label V. The 12 included labels each appear in at least ten percent of the sample articles.

4 Methods

Describe the models/methods you have used including any baselines.

We used three models—logistic regression, CNN, and RNN—for comparison against the pretrained BERT model.

5 Experiments

Describe your experiments. How were the models trained, which hyperparameters, training and inference approaches, etc..

5.1 Results

Report your results using tables and plots as appropriate.

6 Conclusions

Present your discussion/conclusions.

References

Zhiyuan Peng, Behnoush Abdollahi, Min Xie, and Yi Fang. 2021. Multi-label classification of short texts with label correlated recurrent neural networks. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR ’21, page 119–122, New York, NY, USA. Association for Computing Machinery.

Figure 1: Frequency of MeSH Root Labels

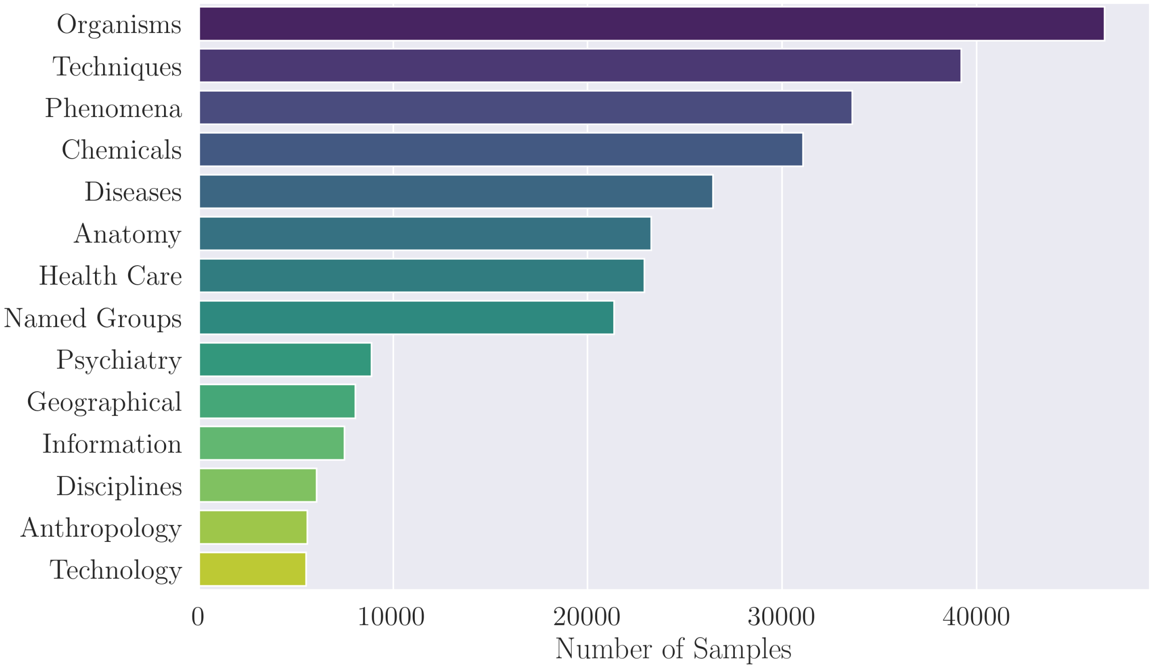

Figure 2: Number of Labels per Abstract

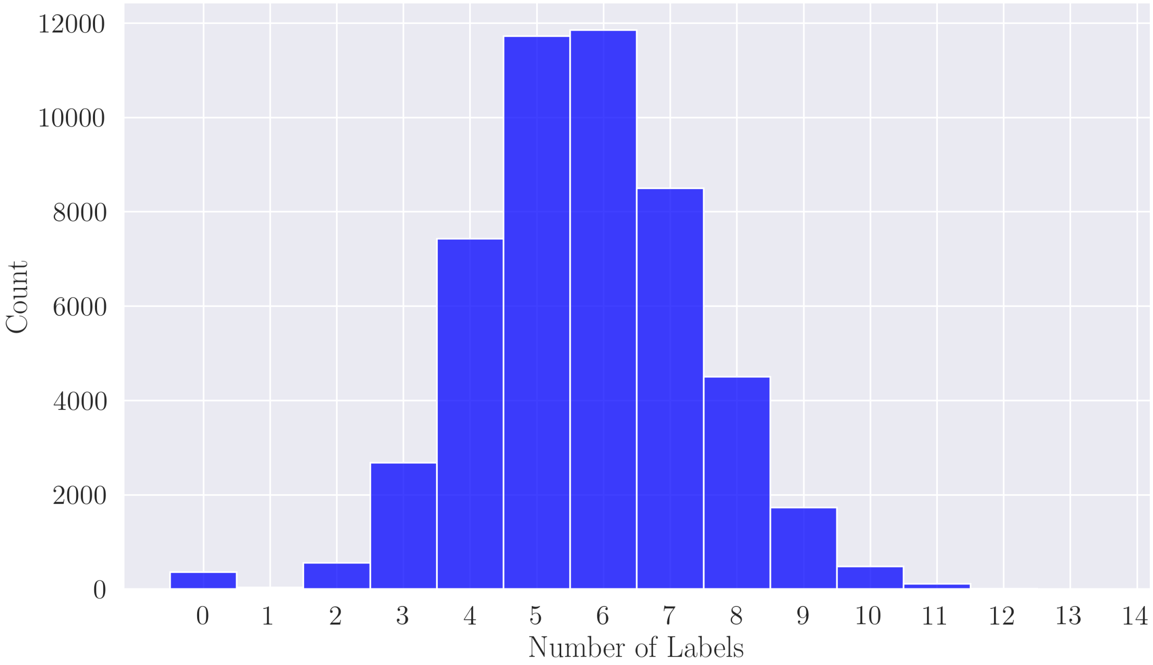

Figure 3: Distribution of Abstract Lengths

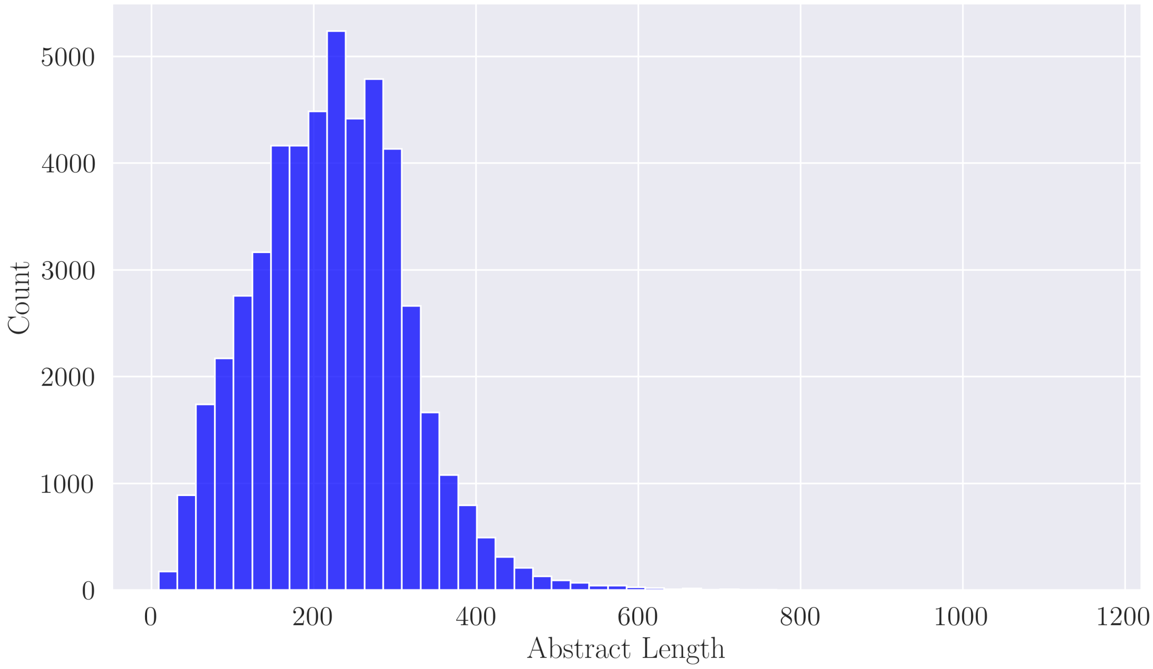

Table 1: Labels

Label	Name
A	Anatomy
B	Organisms
C	Diseases
D	Chemicals and Drugs
E	Analytical, Diagnostic and Therapeutic Techniques, and Equipment
F	Psychiatry and Psychology
G	Phenomena and Processes
H	Disciplines and Occupations
I	Anthropology, Education, Sociology, and Social Phenomena
J	Technology, Industry, and Agriculture
K	Humanities
L	Information Science
M	Named Groups
N	Health Care
V	Publication Characteristics
Z	Geographicals

Table 2: Label Distribution

Label	Train	Dev	Test	All
A	0.4655	0.4629	0.4665	0.4653
B	0.9314	0.9303	0.9330	0.9315
C	0.5280	0.5315	0.5304	0.5291
D	0.6190	0.6225	0.6287	0.6215
E	0.7840	0.7820	0.7858	0.7840
F	0.1768	0.1796	0.1792	0.1777
G	0.6738	0.6640	0.6734	0.6722
H	0.1211	0.1270	0.1177	0.1214
I	0.1116	0.1103	0.1141	0.1119
J	0.1094	0.1109	0.1143	0.1106
L	0.1506	0.1460	0.1515	0.1501
M	0.4242	0.4339	0.4317	0.4273
N	0.4552	0.4628	0.4650	0.4584
Z	0.1610	0.1598	0.1618	0.1610