Document Classification for COVID-19 Literature

Bernal Jiménez Gutiérrez, Juncheng Zeng, Dongdong Zhang, Ping Zhang, Yu Su
The Ohio State University
{jimenezgutierrez.1,zeng.671,zhang.11069,
zhang.10631,su.809}@osu.edu

The global pandemic has made it more important than ever to quickly and accurately retrieve relevant scientific literature for effective consumption by researchers in a wide range of fields. We provide an analysis of several multi-label document classification models on the LitCovid dataset, a growing collection of 23,000 research papers regarding the novel 2019 coronavirus. Additionally, we test these models on a subset of the CORD-19 dataset containing 100 papers about previous epidemics we manually annotated.

Class	LitCovid	CORD-19 Set
Prevention	11,042	12
Treatment	6,897	20
Diagnosis	4,754	25
Mechanism	3,549	70
Case Report	1,914	2
Transmission	1,065	6
General	368	7
Forecasting	461	2

Table 1: Category distribution for the *LitCovid* and *CORD-19 Test Datasets*.

We find that pre-trained language models fine-tuned on this dataset outperform all other baselines and that BioBERT surpasses the others by a small margin with micro-F1 and accuracy scores of around 86% and 75% respectively.

Model	Dev Set		Test Set	
	Acc.	F1	Acc.	F1
LR	68.5	81.4	68.6	81.4
SVM	71.2	83.4	70.7	83.3
LSTM	69.0 ±0.9	83.9 ±0.1	68.9 ±0.3	83.2 ±0.2
LSTM$_{reg}$	71.2 ±0.5	83.9 ±0.3	70.8 ±0.7	83.6 ±0.5
KimCNN	69.9 ±0.2	83.3 ±0.3	68.8 ±0.1	82.7 ±0.1
XML-CNN	72.9 ±0.4	84.1 ±0.2	71.7 ±0.7	83.5 ±0.3
BERT$_{base}$	74.3 ±0.6	85.5 ±0.4	73.6 ±1.0	85.1 ±0.5
BERT$_{large}$	75.1 ±3.9	85.9 ±1.9	74.4 ±2.7	85.3 ±1.4
Longformer	74.4 ±0.8	85.6 ±0.5	73.9 ±0.8	85.5 ±0.5
BioBERT	75.0 ±0.5	86.3 ±0.2	75.2 ±0.7	86.2 ±0.6

Table 2: Performance for each model expressed as *mean* ± *standard deviation* across three training runs.

We evaluate the data efficiency and generalizability of these models as essential features of any system prepared to deal with an urgent situation like the current health crisis.

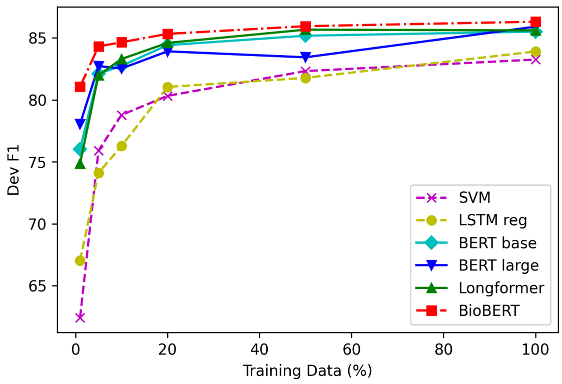

Figure 1: Data efficiency analysis.

All pre-trained language models tested are impressively data efficient, with BioBERT achieving an F1 score only 4 points below its maximum score using only 1% of the training data.

	Acc.	F1
SVM	29.0	62.8
LSTM$_{reg}$	32.7 ±1.5	67.7 ±0.7
Longformer	41.3 ±6.4	70.0 ±2.9
BioBERT	36.0 ±7.8	69.7 ±2.8

Table 3: Performance on the CORD-19 Test Set expressed as *mean* ± *standard deviation* across three training runs.

From Table 3, we can see that performance drops significantly on the CORD-19 test set which does not mention COVID-19. This shows that more work needs to be done for these models to be immediately useful in future health emergencies.

Finally, we explore 50 errors made by the best performing models on LitCovid documents and find that they often (1) correlate certain labels too closely together and (2) fail to focus on discriminative sections of the articles; both of which are important issues to address in future work. Both data and code are available on GitHub [1].

[1] https://github.com/dki-lab/covid19-classification

Acknowledgments

This research was sponsored in part by the Ohio Supercomputer Center (Center, 1987). The authors would also like to thank Lang Li and Tanya Berger-Wolf for helpful discussions.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019a. Docbert: Bert for document classification. *ArXiv*, abs/1904.08398.

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019b. Rethinking complex neural network architectures for document classification. In *NAACL-HLT*.

Simon Baker. 2017. Corpus and Software.

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2016. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32 3:432–40.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Ohio Supercomputer Center. 1987. Ohio supercomputer center.

Q. Chen, A. Allot, and Z. Lu. 2020. Keep up with the latest coronavirus research. *Nature*, 579(7798):193.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Jingcheng Du, Qingyu Chen, Yifan Peng, Yang Xiang, Cui Tao, and Zhiyong Lu. 2019. Ml-net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association : JAMIA*.

Ruihua Fang, Gary Schindelman, Kimberly Van Auken, Jolene Fernandes, Wen J. Chen, Xiaodong Wang, Paul Davis, Mary Ann Tuli, Steven J. Marygold, Gillian H. Millburn, Beverley Matthews, Haiyan Zhang, Nick Brown, William M. Gelbart, and Paul W. Sternberg. 2011. Automatic categorization of diverse experimental information in the bioscience literature. *BMC Bioinformatics*, 13:16 – 16.

Xiangying Jiang, Martin Ringwald, Judith A. Blake, Cecilia N. Arighi, Gongbo Zhang, and Hagit Shatkay. 2019. An effective biomedical document classification scheme in support of biocuration: addressing class imbalance. *Database: The Journal of Biological Databases and Curation*, 2019.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jacob VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

Anthony Rios and Ramakanth Kavuluru. 2015. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. *ACM-BCB ... ... : the ... ACM Conference on Bioinformatics, Computational Biology and Biomedicine. ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, 2015:258–267.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Michael Kinney, Ziyang Liu, William. Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*, abs/2004.10706.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. Sgm: Sequence generation model for multi-label classification. In *COLING*.